POWER AND SAMPLE SIZE DETERMINATION FOR STEPPED WEDGE CLUSTER RANDOMIZED TRIALS

by

Christopher M. Keener

B.S. Molecular Biology, Clarion University of Pittsburgh, 2009

Submitted to the Graduate Faculty of

the Department of Biostatistics

the Graduate School of Public Health in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2018

UNIVERSITY OF PITTSBURGH

GRADUATE SCHOOL OF PUBLIC HEALTH

This dissertation was presented

by

Christopher M. Keener

It was defended on

July 18, 2018

and approved by

Chung-Chou (Joyce) H. Chang, PhD, Professor at Departments of Medicine and Biostatistics School of Medicine and Graduate School of Public Health, University of Pittsburgh

Jesse Yenchih Hsu, PhD, Assistant Professor at Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania

Jong H. Jeong, PhD, Professor at Department of Biostatistics, University of Pittsburgh

Ada O Youk, PhD, Associate Professor at Department of Biostatistics, University of Pittsburgh

John A. Kellum, MD, Professor at Department of Critical Care Medicine University of Pittsburgh

Dissertation Director: Chung-Chou (Joyce) H. Chang, PhD, Professor at Departments of Medicine and Biostatistics School of Medicine and Graduate School of Public Health, University of Pittsburgh

Chung-Chou (Joyce) H. Chang, PhD

POWER AND SAMPLE SIZE DETERMINATION FOR STEPPED WEDGE CLUSTER RANDOMIZED TRIALS

Christopher M. Keener, PhD

University of Pittsburgh, 2018

ABSTRACT

A stepped wedge trial is a type of cluster randomized trial with unidirectional crossover from control to intervention. In this study, we classified stepped wedge trial designs according to subject recruitment and outcome exposure. Based upon those criteria, we proposed three types of classification, that is, fixed cohort (baseline recruitment with longitudinal exposure), expanding cohort (continuous recruitment with longitudinal exposure), and cross-sectional (continuous recruitment with cross-sectional exposure). For each design type, we proposed a corresponding model for estimating treatment effect. We conducted Monte Carlo simula-tions to study the impact of design and analytic assumptions on the sample size and power determination. These assumptions include homogeneous or heterogeneous temporal effects between clusters, fixed or time-varying treatment effect, modeling temporal trend through or not through step effects, and choice of correlation structure. To investigate how these assumptions were made in the published trials, we conducted a systematic review of 300 stepped wedge trials published up to 2017. From the review we found that more than one fourth of these trials did not make it clear in their reports about the type, the assumptions, or models in estimating treatment effect and sample size calculations. The majority of the trials did not mention the methods for handling missing data. This suggests the need for developing standards of reporting stepped wedge trials like CONSORT for randomized trials or STROBE for observational studies.

PUBLIC HEALTH SIGNIFICANCE: Stepped wedge trials are popular for evaluating community-based interventions in public health. This research has focused on three areas of improving the design of a stepped wedge trial: classification of key design aspects, power and sample size determination, and modeling method for estimation and inference the effect of an intervention. Sample size determination is important to ensure that the trial is adequately powered. Model misspecification and incorrect analytic assumption both can lead to inflated Type I error rate or an underpowered trial. Our systematic review found that many stepped wedge trials failed to define key aspects and assumptions of their designs when publishing. Thus, use of our classification of stepped wedge trials will improve technical communication on trials commonly used for public health research.

KEY WORDS : cluster randomized trials, Monte Carlo simulation, power, sample size, stepped wedge trial, systematic review.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1.0 INTRODUCTION

The Stepped Wedge Trial (SWT) Design is a type of cluster randomized trial. A typical parallel cluster randomized trial randomizes clusters (e.g. counties, hospitals, or households) rather than individual to receive either the intervention or the control. Unlike the classic parallel cluster randomized trial, however, in a stepped wedge trial each cluster starts on the control and then crosses over to intervention at some time. Rather than randomize whether a given cluster receives the intervention or the control, in a SWT the time at which the crossover from control to intervention occurs for each cluster that is randomized. The time points at which these crossover events can occur are fixed by a schedule determined at the start of the trial. The time intervals between these potential crossover points are referred to as "steps." These potential crossover times also serve as data collection points. These features of the SWT design can be expressed in a commonly used figure. The figure is essentially a table where the rows refer to clusters (or groups of clusters if more than one cluster is allowed to crossover at a given time) and the columns refer the the discrete time intervals or 'steps.'

Ideally, each cluster has its own unique crossover time. If this is the case, then if the total number of clusters is denoted by I and the total number of steps is denoted by J, then $J = I + 1$. As was mentioned above, it is possible to allow groups of clusters to share the same crossover time. Grouping clusters into the same crossover schedule is often done in order to shorten the total length of the trial. As a consequence, Hussey and Hughes noted that grouping clusters for crossover times results in decreased power due to the resulting lower number of data collection points.

Hussey and Hughes also suggested a general statistical modeling strategy for SWT designs. Their suggested model is a linear mixed model of the form

$$y_{ijk} = \mu + \alpha_i + \beta_j + X_{ij}\theta + \epsilon_{ijk} \tag{1.1}$$

where i refers to the cluster level, j refers to the time point,and k refers to the individual level [10]. The parameters α_i are random effects for the clusters such that $\alpha_i \sim N(0, \tau^2)$. The fixed effects parameters in the model are β_j for the jth cluster and θ is the effect of the intervention. The term ϵ_{ijk} refers to the residual errors such that $\epsilon_{ijk} \sim N(0, \sigma^2)$. As mentioned in their paper, the Hussey and Hughes model can easily be extended into a generalized linear mixed model framework to allow for binary outcomes.

1.0.1 Stepped Wedge Trial Designs from the Published Literature

Copas et al. published a paper that established some important terminology to distinguish between different types of SWT designs [4]. In their paper, they classify SWT designs into one of three categories based on how subjects are enrolled and for how long those same subjects are followed. The three terms they proposed are continuous enrollment with short exposure, a closed cohort, and an open cohort. The continuous enrollment with short exposure type features wholly different individuals in each cluster and step combination. It is commonly referred to as a cross sectional design because each step represents a "snapshot" of that cluster at that point in time. The term "short exposure" is used to note that this trial type is most useful when the nature of the intervention is to provide an effect over a small period of time.

Closed cohort trial types select a group of subjects at the start of the trial to follow throughout the course of the study. Thus, compared to the cross sectional design type, a closed cohort should additionally account for correlation between measurements on the same individual over time. In a paper published by Baio et al [2], the model 1.1 was modified by the addition of a random effect for subject ν_{ik} to account for correlation of measurements on the same subject.

$$\phi_{ij} = g(\mu_{ij}) = \mu + \alpha_i + \beta_j + X_{ij}\theta + \nu_{ik} \tag{1.2}$$

An open cohort trial type starts with the enrollment of a large group of subjects at the start of the trial also. However, some subjects might be allowed to enroll later on or drop out of the study early. Copas et al even suggested that individuals might switch between clusters in an open cohort trial type.

1.0.2 Power and Sample Size Estimation from the Published Literature

Hussey and Hughes[10] provided an analytical formula for sample size and power when all the variance components were known. They also noted that treatment delays result in reduced power. In all cases, their methods focused only on the *cross-sectional* SWT design. Woertman et al.[19] provided an alternative to calculating sample size and power for cross-sectional SWT designs through a design effect. They assumed homogeneous temporal effects across clusters and that the same number of clusters cross over for each step. The design effect they provided can be used to estimate sample size by first computing the sample size for a subject-level randomized controlled trial, then multiplying this sample size by the design effect. Hemming et al.[6] defined incomplete SWT designs and provided sample size and power calculations for incomplete cross-sectional designs. An incomplete SWT design is one in which outcomes are not measured for every cluster at every step. Additionally, they provided sample size and power estimation for designs with multiple levels of clustering due to groups within the randomization-level clusters. Their sample size and power estimation was based upon mixed models in which the variance components are known. Baio et al. [2]

suggested that an analysis model for closed cohorts should incorporate a random subject effect. They outlined a general simulation procedure for estimating sample size and power for SWTs. They compared their power and sample size results from the analytical formula estimates from Hussey and Hughes [10] for cross-sectional designs. They compared the sample size estimates using the analytical power formula from Hussey and Hughes[10] and the design effect method from Woertman et al.[19]. They also provided power curves for a closed cohort design based upon Monte Carlo simulations. Hooper et al.[8] published a paper that offered sample size calculations for cross-sectional and closed cohort SWT design computed using a design effect. For this computation, a design effect for clustering is computed as well as a design effect for the repeated measures cross-sectionally taken for each cluster.

1.0.3 Author Contributions to the Literature

My work extends power and sample size considerations for stepped wedge cluster randomized trials. To do so, I first proposed changes to terminology classifying SWT designs. My terminology focuses on two main features of SWT design: subject recruitment and duration of subject inclusion in the trial. From these two features, I proposed the types of stepped wedge trial designs as cross-sectional, fixed cohort, and expanding cohort. I then describe a number of design and analytical assumptions that must be understood about SWT designs to obtain power estimates, accurately estimate the intervention effect, and make proper statistical inference regarding the effect of intervention. These assumptions are homogeneity versus heterogeneity of temporal trends, time-varying versus fixed treatment effects, adjustment for cluster-level and subject-level temporal trends, choice of correlation structure, and choice of fitting the step effect as fixed versus random.

To assess the quality of general design reporting as well as power and sample size estimation, I conducted a systematic review of 300 published SWTs through 2017. For each of these trials, I reported the design type, homogeneity versus heterogeneity of temporal trends, time-varying versus fixed treatment effects, method used to adjust for temporal trends, choice of correlation structure, choice of fitting the step effect as fixed versus random, and method

used for missing data. Grayling et al.[5] previously published a review of SWTs. While this review was very informative, my review additionally includes 2016 and 2017. Moreover, my review focused on the aforementioned design and analytical issues, which were not considered in the review by Grayling et al.[5].

For the aforementioned design and analytical issues, I conducted Monte Carlo simulation studies. Each of these studies were conducted for each of the three SWT design types: cross-sectional, fixed cohorts, and expanding cohorts. In addition to reporting power under various design and analytical circumstances, I also studied how analysis model misspecification affects the estimation and inference for the effect of intervention. For heterogeneity of temporal trends, I consider a misspecified model with homogeneous temporal trends. For time-varying intervention effect, I consider a misspecified model with time-fixed intervention effect. For choice of correlation structure, I consider data generated with exchangeable and autoregressive structures and analysis models using each structure as well. For the topic of fitting steps as random effects, I consider the case where the step effects are actually fixed effects as well as the case when there is an interaction between step and the effect of intervention. Specifically just for expanding cohort designs, I conducted a Monte Carlo simulation study to determine failure to account for both cluster-level and subject-level temporal trends might influence the estimation or inference for the effect of intervention. For each simulation study the bias, average model standard error, Monte Carlo standard error, coverage probability, and power are reported.

The remainder of this document has the following structure. Chapter two established my terminology for SWT design types and discusses the analytical and design issues. This chapter also includes the systematic review of 300 published SWTs up to 2017. Chapter two concludes with a brief description of an R package I am publishing for power and sample size estimation for SWT designs with the various discussed features. Chapter three covers the simulations studies conducted for the design assumptions including homogeneity versus heterogeneity of temporal trends, time-varying versus time-fixed effect of intervention, and fitting both cluster-level and subject-level temporal trends for expanding cohort designs. Chapter three contains the simulation studies for analytical issues including choice of correlation structure and fitting the steps as random effects.

2.0 SAMPLE SIZE AND POWER DETERMINATION FOR STEPPED WEDGE CLUSTER RANDOMIZED TRIALS

ABSTRACT

The stepped wedge trial is a type of cluster randomized trial with unidirectional crossover from control to intervention. We discuss a number of design and analytical issues that relate to accurate power and sample size estimation. We reviewed existing literature for power and sample size estimation for stepped wedge trials including topics such as analytical estimations, treatment delays, design effects, incomplete designs, and simulations. To facilitate power and sample size discussion for the various stepped wedge design types, we proposed new terminology to classify stepped wedge trials according to subject recruitment and outcome exposure: fixed cohort (baseline recruitment with longitudinal exposure), expanding cohort (continuous recruitment with longitudinal exposure), and cross-sectional (continuous recruitment with cross-sectional exposure). For each of these three design types, we recommended analysis models for estimating the effect of intervention.

We conducted a systematic review of 300 stepped wedge trials published up to 2017. For this review, we focused on our three stepped wedge trial types and analysis features important to accurate power and sample size estimation such as heterogeneity of temporal effects, fixed versus time-varying treatment effects, adjustments for time effects, choice of correlation structure, choice to fit the step effect as fixed or random, and methods for analyzing missing data. The literature review found that one fourth or more of trials failed to clearly report these details in the design and analytical aspects and that the majority of trials failed to describe how missing values were handled in the analyses. Our review demonstrates the importance of development and adherence to a set of reporting guidelines for stepped wedge

trials similar to the CONSORT guidelines for randomized trials or STROBE for observational studies.

Stepped wedge trials are popular for evaluating community-based interventions in public health. For that reason, ensuring stepped wedge trials are adequately powered is important to public health research. The systematic review we conducted discovered that many stepped wedge trials did not clearly define key aspects and assumptions of their designs when publishing. These aspects are important to ensure that power and sample size estimation is accurate. Thus, use of our terminology for stepped wedge trials as well as reporting those design and analytical assumptions that we discussed will improve technical communication on stepped wedge trials for public health research.

KEY WORDS : Cluster randomized trial, power, sample size, stepped wedge trial, systematic review.

2.1 INTRODUCTION

The stepped wedge trial (SWT) is a type of cluster randomized controlled trial in which the unit of randomization is cluster (e.g., clinic), not subject (e.g., patients having care in the clinic). At the start of a SWT, all clusters begin with control groups. At certain time points during the course of the trial, some of the clusters are randomly selected to be switched to the intervention/treatment groups. The time points for clusters to change status from control to intervention usually are evenly spaced and are determined before the start of the trial. These blocks of time between such preset crossover times are referred to as *steps*. Although a variety of SWT designs exist, a general representation of a SWT design can be depicted in Figure 1.

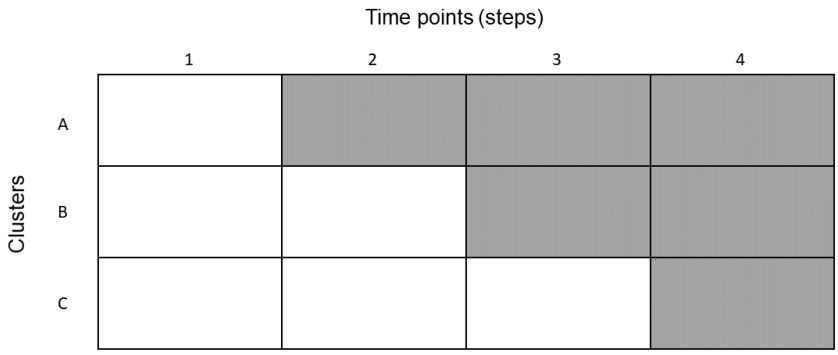

Figure 1: An illustration of a typical intervention schedule for a SWT design.

To determine the needed sample size or power for designing a SWT, we must elucidate the associated design and analytical assumptions. After reviewing and examining the designs of 300 studies (Supplement A) for the SWTs published from 1987 to 2017[5], we found that many of these studies did not clearly specify all the necessary assumptions and analytical parameters for their trials, therefore, raise doubts about the appropriateness of their proposed analytic methods, estimations, inferences, sample size and power calculations, and even the findings. Because there is no prior published work clearly specifies various types

of SWT designs and their associated assumptions, in this study we filled in this gap by defining different types of SWT designs and identifying the associated design and analytical assumptions in each of the design type.

In addition, we developed a program in R to estimate sample size and power for each type of the designs properly using a framework of generalized linear mixed effects models (GLMMs) with the associated design assumptions and user-selected analytical assumptions.

The paper is organized as follows. In Section 2.2, we describe and define different types of SWT designs. In Section 2.3, we summarize methods for sample size and power calculations in the published literature. In Section 2.4, we discuss design assumptions that will impact the estimation of sample size and power. In Section 2.5, we present the results of our systematic review of stepped wedge trials between 1987 and 2017. In Section 2.6, we provide two examples of actual SWTs to demonstrate sample size and power related issues via simulation. In Section 7, we discuss an R program that will be released to compute sample size and power for stepped wedge trials.

2.2 TYPES OF SWT DESIGNS

Copas et al.[4] classified SWTs into three main types of designs: continuous recruitment with short exposure (CRSE), closed cohorts, and open cohorts. For the CRSE design, subjects are only exposed for a short period of time. For an intervention deployed in intensive care units (ICUs), for instance, subjects' exposure period is their ICU length of stay. Since the exposure period is short, these subjects can only serve as either the control or the intervention but not both. Recruitment continues throughout the duration of the trial as new subjects become exposed. In a closed cohort design, all subjects are enrolled at the beginning of the trial. Subjects remain exposed and are followed for the duration of the trial. A trial with an open cohort design involves a long period of exposure as that in a closed cohort trial. However, unlike a closed cohort design, the trial with an open cohort design allows new subjects to be recruited after the study starts. There are two main subtypes to the open cohort design:

one is similar to that of the closed cohort design except that late enrollment and dropout are allowed; another subtype measures outcomes only on a random sample of the recruited subjects. For example, investigators might want to design a cross-sectional SWT to study a community-wide public health intervention such as an initiative for cleaner water. Such a community-wide clean water initiative would have long lasting effects rather than a short period of exposure such as for an ICU-based study.

While the classification proposed by Copas et al.[4] is useful in many ways there possible improvements. For a trial with a closed cohort design, it is not useful to stipulate the need for complete follow-up when in many cases that cannot be guaranteed by the design. Moreover, the two subtypes for the open cohort designs differ in basic design characteristics, namely whether subjects are longitudinally followed through steps or not. In the open cohort designs with longitudinal follow-up, sample size estimation should take into consideration on how many of the cohort are enrolled at the start of the study, and at what rate new subjects are enrolled afterward. In the open cohort design subtype based on random samples, sample size is based on consistent sampling throughout the study. However, specification of an analysis model depends on when subjects are recruited and the frequency and manner of outcome measurements. We describe and define different timings of recruitment and outcome measurement, also propose a different classification method for SWT designs based upon subject recruitment and subject exposure.

If all subjects were recruited at the beginning of the trial, we define the design of such a trial as having *baseline recruitment*. Conversely, if subjects may be recruited over the course of a trial, we define the design of this type of a trial as having *continuous recruitment* SWT.

Regarding outcome exposure, if subjects are potentially followed for more than one step, we define the design of such a trial as having *longitudinal exposure*. One example of this type of design is when repeated measurements are taken on the same subject over multiple steps as they remain in the trial. Other examples include trials with time-to-event outcomes and incident count (or rate) outcomes. Conversely, if each subject is only followed for one step, we define this type of a trial as having *cross-sectional exposure*.

By combining the characteristics of recruitment and outcome exposure, we propose three types of SWT designs: *fixed cohorts* (baseline recruitment with longitudinal exposure), *expanding cohort* (continuous recruitment with longitudinal exposure), and *cross-sectional* (continuous recruitment with cross-sectional exposure). Note a SWT design cannot have *baseline recruitment* and *cross-sectional exposure* in one trial.

A *fixed cohorts* SWT design is essentially the *closed cohort* trial in Copas et al.[4] One difference between the term *fixed cohort* in our definition and *closed cohort* in Copas et al.[4] is that the *fixed cohort* allows for subject withdrawal since withdrawal cannot be prevented by design alone in most cases. Excluding withdrawal, each subject experiences both the control and intervention with the timing of crossover being randomized by clusters. Figure 2 depicts how a *fixed cohort* SWT operates. Equation (1) is a generalized linear model with link function $g(\cdot)$ for analyzing fixed cohort SWT designs.

$$g(y_{k(i)j}) = \mu + u_i + u_{k(i)} + \beta_j + \theta trt_{ij} \tag{2.1}$$

$$u_i \sim N(0, \sigma^2_{cl})$$

$$u_{k(i)} \sim N(0, \sigma^2_{subj})$$

where i, j, k, μ, u_i, $u_{k(i)}$, β_j, θ, trt_{ij}, σ^2_{cl},σ^2_{subj} are the cluster index, the step index, the subject index, the mean outcome for time zero on control, the random effect for each cluster, the random effect for each subject, indicators variables for the time effects, the fixed effect of intervention, an indicator for whether cluster i is on intervention at step j, the variance for the random cluster intercepts, and the variance component for the random subject intercepts.

An *expanding cohort* SWT design is similar to a *fixed cohort* design except that it allows for the enrollment of additional subjects during the trial. Figure 3 depicts the *expanding cohort* SWT design. An expanding cohort design can be analyzed using the same model as for *fixed cohort* designs (see Equation 2.1).

Trials with *cross-sectional* design do not begin by recruiting a cohort of subjects. There are two main types of *cross-sectional* SWT designs based on the length of follow up. If the follow-up period is short, the study fits the term *continuous recruitment, short enrollment*

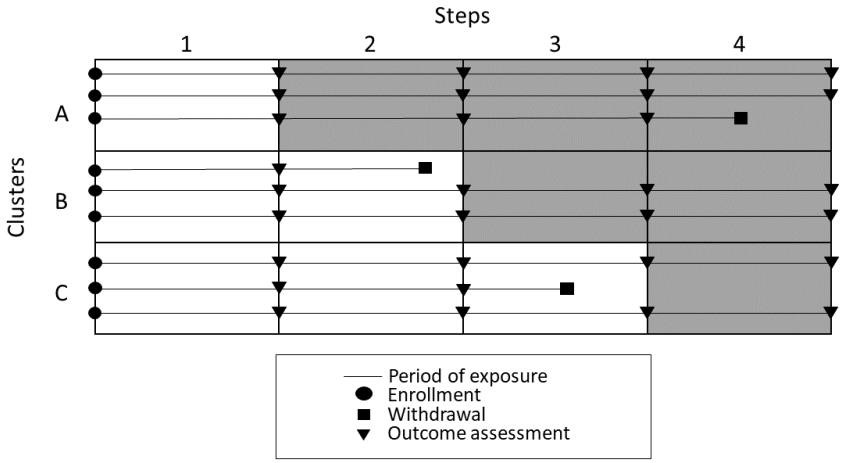

Figure 2: A diagram demonstrating baseline recruitment and longitudinal exposure features of a *fixed cohort* SWT design.

Note that the unshaded regions indicate control, and the shaded regions indicate intervention.

in Copas et al.[4] Conversely, if the exposure time is long, then by the terminology in Copas et al., we could term them as an *open cohort study* with *cross-sectional sampling*. Figures 4 and 5 illustrate the two main subtypes of *cross-sectional* SWT designs. Equation 2.2 is a generalized linear model with link function $g(\cdot)$ for analyzing *cross-sectional* SWT designs.

$$g(y_{k(ij)}) = \mu + u_{0i} + \beta_j + \theta trt_{ij} \tag{2.2}$$

$$u_{0i} \sim N(0, \sigma^2_{cl})$$

In Equation 2.2, μ is the mean outcome for time zero on the control. The parameters β_j are indicator variables for the time point (i.e. step). The parameter θ is the fixed effect for the intervention. The random effects are u_{0i} the random cluster intercepts.

In some cases, a SWT does not collect outcome data for every cluster at every step. Such trials are referred to as incomplete stepped wedge trials. *Fixed cohort*, *expanding cohort*, and *cross-sectional* trials can also be incomplete designs. Under the model specification in Equations (2.1) and (2.2), incomplete designs can be analyzed like a complete design with

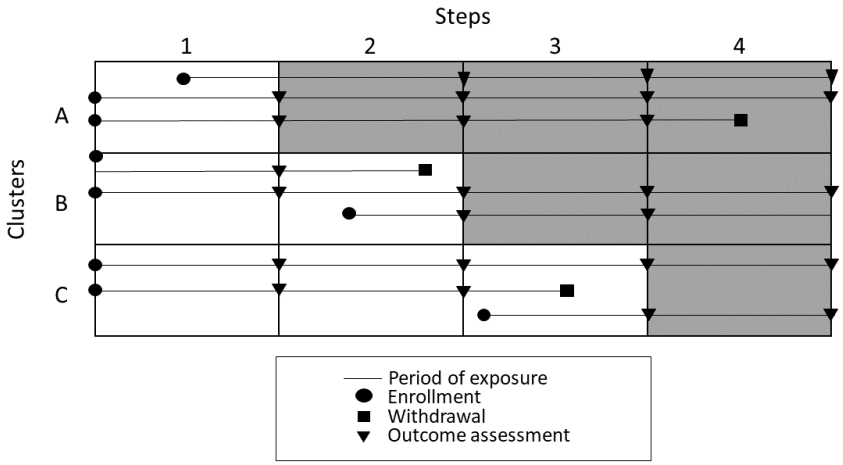

Figure 3: A diagram demonstrating continuous recruitment and longitudinal exposure features of an *expanding cohort* SWT design.
Note that the unshaded regions indicate control, and the shaded regions indicate intervention.

only a minor loss of power due to having less data since mixed models only require data to be missing at random (MAR). However, except for incomplete designs with balanced missing values (i.e., the same step(s) missing for every cluster and the subjects nested within those clusters), generalized estimating equations (GEE) cannot be used for the analysis.

2.3 SAMPLE SIZE AND POWER IN THE CURRENT LITERATURE

Hussey and Hughes[10] suggested the use of mixed effects models for estimating treatment effect and provided an analytical formula for sample size and power when all the variance components were known. The effects that treatment delays have on power were also discussed. In all cases, their methods focused only on the *cross-sectional* SWT design.

Woertman et al.[19] provided an alternative to calculating sample size and power through a design effect. Their method assumes no within-subject correlation over time, which is appropriate for *cross-sectional* SWTs without longitudinal measures since in such designs

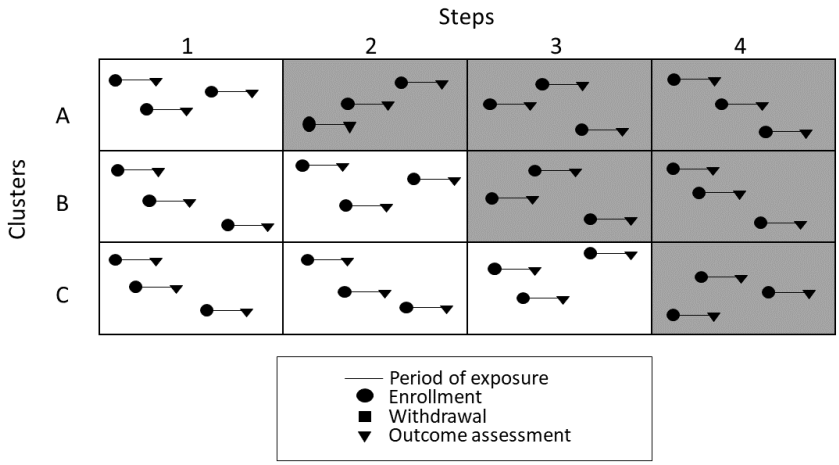

Figure 4: A diagram demonstrating continuous recruitment and cross-sectional exposure features of a *cross-sectional* SWT design with short exposure period and continuous recruitment.

Note that the unshaded regions indicate control, and the shaded regions indicate intervention.

each subject has only one measurement. They also assumed homogeneous temporal effects across clusters and that the same number of clusters cross over for each step. The design effect they provided can be used to estimate sample size by first computing the sample size for a subject-level randomized controlled trial, then multiplying this sample size by the design effect.

Hemming et al.[6] defined an incomplete SWT design and provided sample size and power calculations for cross-sectional designs with these features. An incomplete SWT design is one in which outcomes are not measured for every cluster at every step. They also provided sample size and power estimation for designs with multiple levels of clustering due to the existence of groups within the randomization-level clusters. Their sample size and power estimation was based upon mixed models in which the variance components are known.

Baio et al. [2] suggested that analysis model for a SWT design with repeated measures on a subject should incorporate a random subject effect. They outlined a general simulation procedure for estimating sample size and power for SWTs. They compared their power

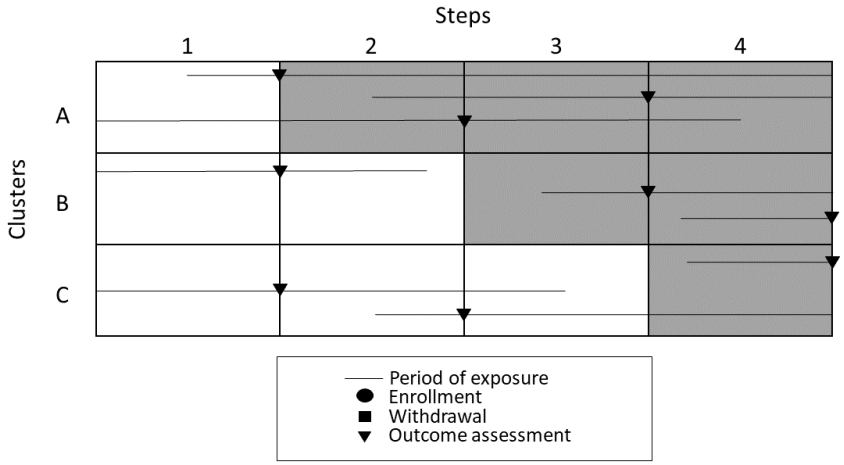

Figure 5: A diagram demonstrating continuous recruitment and longitudinal exposure features of a *cross-sectional* SWT design with recruited subjects as a sample of the exposed population of the study.
Note that the unshaded regions indicate control, and the shaded regions indicate intervention.

and sample size results from the analytical formula estimates from Hussey and Hughes [10] for cross-sectional designs. They compared the sample size estimates using the analytical power formula from Hussey and Hughes[10] and the design effect method from Woertman et al.[19]. They also provided power curves for a closed cohort design based upon Monte Carlo simulations.

Hooper et al.[8] published a paper in which they offered sample size calculations for cross-sectional and closed cohort SWT design computed using a design effect. For this computation, a design effect for clustering is computed as well as a design effect for the repeated measures cross-sectionally taken for each cluster.

While these publications provide useful information on sample size and power for certain types of SWT designs, the current literature lacks a systematic evaluation of sample size and power for different types of SWT designs under various assumptions made.

2.4 DESIGN ASSUMPTIONS

There are three important design assumptions that are necessary to study sample size and power for SWT designs. The three assumptions are: (1) temporal (i.e., step) effects are homogeneous or heterogeneous between clusters; (2) treatment effect is fixed or time-varying; and (3) existence of cluster-level and subject-level temporal trends.

2.4.1 Temporal effects are homogeneous or heterogeneous between clusters

The published SWT literature has stated the importance of adjusting for background temporal effects. When analyzing SWTs researchers often assumed that the background temporal effects are homogeneous across clusters. Hussey and Hughes [10] briefly discussed the possibility of heterogeneity of temporal effects across clusters. In that paper, the temporal (i.e., step) effect was fitted as a series of indicator variables to allow for nonlinear temporal effects. The authors noted that fitting an interaction term between the step indicators variables and the cluster effects would result in an overparameterized model at the cluster level. For that reason, Hussey and Hughes suggested grouping clusters with similar characteristics together and fitting interaction terms between those groups and the time. When temporal trend effects are homogeneous across clusters, the typical model specification can be used such as Equation 2.2 or Equation 2.1 for cross-sectional and for cohort designs respectively. When the temporal trends might be heterogeneous across clusters, we recommend adding random time slopes for each cluster in Equations (2.1) or (2.2). For example, for a *cross-sectional* design, Equation (2.2) becomes Equation (2.3).

$$g(y_{k(ij)}) = (\mu + u_i) + \theta trt_{ij} + (\beta + b_i)t_j \tag{2.3}$$

for subject k in cluster i at time j where

$$\begin{bmatrix} u_i \\ b_i \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2_{cl} & \sigma_{cl,step} \\ \sigma_{cl,step} & \sigma^2_{step} \end{bmatrix} \right).$$

Similarly, for either a *fixed cohort* or an *expanding cohort* design Equation (2.1) becomes Equation (2.4).

$$g(y_{k(ij)}) = (\mu + u_i + u_{k(i)}) + \theta trt_{ij} + (\beta + b_i)t_j \tag{2.4}$$

for subject k in cluster i at time j where

$$\begin{bmatrix} u_i \\ b_i \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2_{cl} & \sigma_{cl,step} \\ \sigma_{cl,step} & \sigma^2_{step} \end{bmatrix} \right).$$

In Equation (2.4), u_i, b_i, σ^2_{cl}, σ^2_{subj}, and $\sigma_{cl,subj}$ are the random cluster intercept, the random slope on time for each cluster, the variance of the random intercepts, the variance of the random slopes, and the covariance between the random intercepts and random slopes.

2.4.2 Treatment effect is fixed or time-varying

Most analytical models for SWTs consider the effect of the intervention to be time-invariant. Time-fixed treatment effects can be seen in the basic suggested analysis models in Equations (2.2) and (2.1). However, there are certain situations in which one might expect the effect of intervention to change over time. For example, the research staff might become more adept at delivering the intervention over time.

Hussey and Hughes[10] noted that the effect of the intervention might not be fully realized during a single step. They suggested that in the middle of the study, we may use the design effect for intervention which is a factional value of treatment indicator (i.e., between 0 and 1) in place of the intervention design effect X_{ij} (with a value of 1) to reflect the incomplete achievement of full intervention effect. This topic was later revisited in another paper by Hughes et al. [9] that provided a more detailed on the use of fractional X_{ij} values and the effects on power. They provided a formula for the estimated treatment effect parameter when a treatment delay exits but is not accounted for in the analysis. The major limitation to this strategy is that an analyst must provide the fractional values for X_{ij} rather than estimating them from the data.

Hughes et al.[9] provided another more data-driven approach for time-varying treatment effects. They suggested a method for explicitly parameterizing and estimating the time-on-treatment effect. The procedure starts by fitting the following model

$$Y_{ijk} = \mu + \alpha_i + \beta_j + L_{ijl}\theta_l + e_{ijk}, \tag{2.5}$$

where $l = \sum_{r=1}^{j} X_{ir}$ is the number of time steps since the introduction of the intervention. Then, L_{ijl} is an indicator variable for whether cluster i has been in the intervention for l steps by time step j. Alternatively, a linear parametric form for the time-on-treatment effect was suggested such as $L_{ijl} = l$ and $\theta_l = \theta$. This model is particularly appropriate when the investigators believe that the intervention effect is increasing over implementation time to a stable long term effect, such as for a delay in the effect or implementation of the intervention.

One might imagine, on the other hand, that the effect of intervention might have a synergism with temporal trends. In such cases, the model can include an interaction term between the intervention effect and time such as shown in Equation (2.6) for a *cross-sectional* design, where γ represents the parameter for a time-varying treatment effect.

$$g(y_{k(ij)}) = \mu + u_i + \theta trt_{ij} + \beta t_j + \gamma trt_{ij} * t_j \tag{2.6}$$

2.4.3 Cluster-level vs. Subject-level Temporal Trends

As mentioned in Section 2.1, it has been established in the stepped wedge literature that adjusting for background temporal trends is important for obtaining an unbiased estimate of the treatment effect. For *cross-sectional* SWT designs, since there are no repeated measures for subjects over the course of the trial, we only need to consider cluster-level time trends. A cluster-level time trend might include changes in policy or seasonal fluctuations. On the other hand, cohort (*fixed* or *expanding*) SWT designs involve repeated measures on subjects throughout the course of the trial. Subject-level time trends might represent the natural history of the condition being studied. Analyses should adjust for subject-level time trends to prevent them from confounding the effect of intervention.

For *fixed cohort* SWT designs, all subjects are enrolled at the start of the trial. If we design, for example, a SWT to study a population afflicted by a certain condition, having a *fixed cohort* would mean that we select a cohort that all has the condition at the onset of the trial. Thus, the subject-level time trend for the course of the condition would be on a set time frame with any cluster-level time trends. Conversely, for *expanding cohort* SWT designs, however, subjects can be recruited later in the study. Some subjects might acquire the condition and be recruited in the middle of the study. With some subjects acquiring the condition of interest after the first step, subject-level temporal trends are no longer on a set time frame along with the cluster-level temporal trend. Thus, we argue that for *expanding cohort* SWT designs, it is appropriate to adjust for both cluster-level and subject-level temporal trends.

2.5 ANALYTICAL ASSUMPTIONS

In the same manner that SWTs can vary greatly in terms of the assumption in the design, there are also assumptions in how the SWT data are analyzed. To estimate sample size correctly, it is important to consider the user-defined analytical assumptions. In this section, we shall explore two different aspects for analysis of SWTs: (1) fitting the step effect as a fixed or a random effect and (2) the choice of correlation structure.

2.5.1 Fixed versus random effects for step

In Hussey and Hughes [10], they recommended using indicator variables for the step in order to account for secular trends. The authors made a brief comment about a suggestion from reviewers for whether the step effect could be fit as random effects instead. They replied that "We felt that this approach did not reflect our interest in controlling for temporal trends and fluctuations in disease prevalence over the course of a particular trial" [10]. Nonetheless, they did suggest that random effects for step might be appropriate in other settings and is worth further investigation.

Random effects are suggested when the effects in question represent a random subset of a larger set of effects. From this perspective, the steps in a SWT seem more appropriately fitted as random effects. To clarify the statements made by Hussey and Hughes [10], however, the most common aim of modeling step effects is to control for temporal trends in the outcome. Thus, the steps effects really represent a series of indicator variables for a nonparametric time trend. If the investigators suspect that there is random variation among the steps not due to temporal trends (but rather due to issues such as staff turnover), random effects could be useful. However, it is still important to ensure that any temporal trends are properly adjusted in the model so that the intervention effect (with unidirectional crossover) is not confounded by these temporal trends.

One model recommendation in such a case would be to model both a parametric term for the temporal trend as well as random effects for the steps as in Equation (2.7) for a *cross-sectional* design or Equation (2.8) for *fixed cohort* or *expanding cohort* designs.

$$g(y_{k(ij)}) = \mu + u_i + \theta trt_{ij} + \beta t_j + b_j \tag{2.7}$$

such that

$$u_i \sim N(0, \sigma^2_{cl})$$
$$b_j \sim N(0, \sigma^2_{step})$$

$$g(y_{k(ij)}) = \mu + u_i + u_{k(i)} + \theta trt_{ij} + \beta t_j + b_j \tag{2.8}$$

such that

$$u_i \sim N(0, \sigma^2_{cl})$$
$$u_{k(i)} \sim N(0, \sigma^2_{subj})$$
$$b_j \sim N(0, \sigma^2_{step})$$

For both Equations (2.7) and (2.8) β is the fixed effect parameter for the temporal trend, t_j are the time points, and b_j are the random effects for the steps.

2.5.2 Choice of correlation structure

Hussey and Hughes [10] mostly focused on cross-sectional SWT designs, and suggested fitting a random effect for cluster in a linear mixed model (LMM) to account for correlation in outcomes from subjects within the same cluster. One drawback to LMMs is that their results are sensitive to the choice of correlation structure. Conversely, GLMs fit with GEE are known for providing results that are robust to misspecification of the working correlation structure. [13] The drawback to using GEE is that the standard empirical variance estimator relies on asymptotic properties. Thus, a GEE-based analysis of a SWT with few clusters might yield inaccurate inference. Scott et al. [18] studied the use of various finite-sample corrected variance estimates for GEE in SWT designs. Despite these facts, LMM remains the most popular methods for analyzing SWTs.

Kasza et al.[12] suggested another correlation structure for cross-sectional SWT designs. They suggested use of non-uniform correlation structure so that not all pairs of measures within the same cluster would share the same correlation. Their suggested correlation structure ensures that measures within the same cluster and step are exchangeable, but measures that are in the same cluster but in different steps can have decaying correlation based upon the difference in time.

When fitting LMMs or GLMMs for SWT data, investigators should examine different potential correlation structures before arriving at the final model. For covariance pattern models with a normally distributed outcome, the model is based on likelihood, so information criteria such as Akaike's Information Criteria (AIC)[1] or Bayesian Information Criteria (BIC)[17] can be used for model selection. If the two correlations structures are nested models, a likelihood ratio test can be used. Covariance pattern models with non-normal outcomes can be estimated from psuedo-likelihood. When fitting such models using SAS's PROC GLIMMIX, users can use Pseudo AIC and Pseudo BIC to choose the correlation structure[11], though using these metrics for choosing fixed or random effects is not advised. Despite the fact that misspecification of the working correlation structure does not bias estimates for the mean model[13], correct specification will lead to greater efficiency. One can compare different correlation structures within GEE based upon quasi-likelihood using QIC[16].

2.6 SYSTEMATIC LITERATURE REVIEW

Grayling et al.[5] published a systematic review of protocols and manuscripts for SWTs published on or before 2015. Their search was conducted across PubMed (including Medline), Ovid (including Embase), Web of Knowledge, PsycINFO, the Cochrane Library, the ISRCTN registry, and ClinicalTrial.gov. Search terms included 'stepped wedge,' 'step wedge,' 'experimentally staged introduction,' 'delayed intervention,' and 'one directional crossover design.' To qualify as a SWT design, each study met the following criteria: cluster-level allocation of the intervention, at least three steps, at least two clusters, unidirectional crossover from control to intervention, time of crossover randomized, complete SWT design, and data collected for a step in which all clusters were on intervention. The exact protocol for the review is provided by Grayling et al.[5]

We extended the review to the end of 2017. The term 'stepped wedge' was queried for those years in the PubMed database. Trial identification numbers (specifically ClinicalTrials.gov, Netherlands Tiral Register, ISRCTN registry, Australian New Zealand Clinical Trials Registry, and the Pan-African Clinical Trials Network registry) were extracted from publications when available in order to link together publications from the same trial. Figure 6 shows the process of classifying publications as from the same trial for the 300 trials included in this review.

The results of the systematic literature review are presented in Table 1. For design type, 170 (56.67%) of the 300 studies used an *cross-sectional design*. *Fixed cohort* designs accounted for 79 (26.33%), and *expanding cohort* designs accounted for 32 (10.67%). Four (1.33%) of the trials used a design that could not be assigned a traditional stepped wedge design type, and 15 (5.00%) did not provide adequate information on design to classify them.

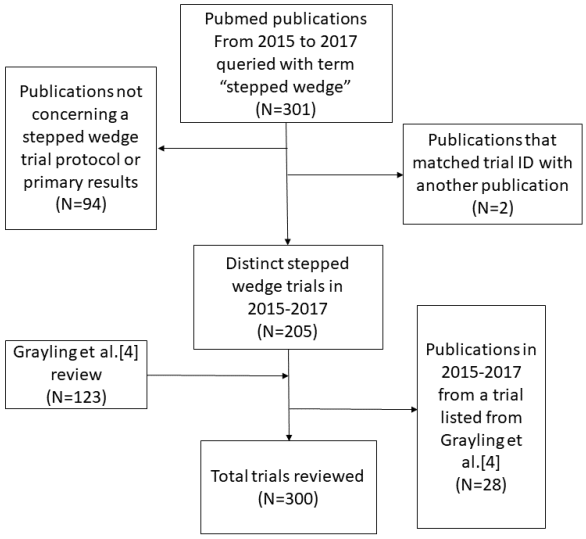

Figure 6: This flowchart demonstrates how publications were combined into the review while avoiding multiple counts for trials with more than one publication.

We also review the design assumptions made in these designs. In terms of homogeneity or heterogeneity of temporal effects, 116 (38.67%) did not clearly state whether the temporal effect was homogeneous or not. One hundred forty-six (48.67%) studies specified that the temporal effect was homogeneous. Nineteen (6.33%) trials did not use temporal effects at all. Nineteen (6.33%) studies indicated that the temporal effects varied, with 5 (1.67%) of those being heterogeneous by covariate.

For treatment effect, 178 (59.33%) trials fitted treatment effect as fixed, and 83 (27.67%) did not specify. Thirty-six (12.0%) studies incorporated a time-varying treatment effect. One (0.33%) study let the treatment effect vary by cluster.

Time was accounted for through regression analysis in 160 (53.33%) trials. Twenty-seven (9.00%) studies did not adjust for temporal effects. One hundred one (33.67%) trials did not clearly indicate whether temporal effects were accounted for in the analysis.

Table 1: The results of our systematic review of SWTs published between 1987 and 2017 are presented in this table.

		Frequency	Percentage
Trial Design Type	Closed longitudinal	79	26.33%
	Cross-sectional	170	56.67%
	Open longitudinal	32	10.67%
	Other	4	1.33%
	Unclear	15	5.00%
Temporal Effects	Heterogeneous by covariate	5	1.67%
	Heterogeneous by cluster	14	4.67%
	Homogeneous	146	48.67%
	None	19	6.33%
	Not stated	116	38.67%
Correlation structure			
	For Both Clusters & Subjects		
	clusters, subjects: random effects, AR	4	1.33%
	clusters, subjects: random effects, compound symmetric	1	0.33%
	clusters, subjects: random effects, random effects	30	10.00%
	clusters, subjects: random effects, random slopes	1	0.33%
	clusters, subjects: random effects, unclear	1	0.33%
	For Clusters Only		
	Boot-strapped standard errors	1	0.33%
	Clustering standard errors	1	0.33%
	Gamma frailty	1	0.33%
	GEE	16	5.33%
	Random effects	102	34.00%
	Robust standard errors	1	0.33%
	For Subjects Only		
	AR	2	0.67%
	Random effects	9	3.00%
	Independent	30	10.00%
	Not stated	85	28.33%
	Random effects (level unclear)	5	1.67%
	Unclear	4	1.33%
	Wilcoxon Signed Rank	1	0.33%
	Other*	4	1.33%

Table 1: continued

Methods for missing data			
	MCAR assumed	1	0.33%
	Complete case analysis	21	7.00%
	Imputation	43	14.33%
	Best-subset imputation	1	0.33%
	Fixed imputation	7	2.33%
	Last value carry forward imputation	2	0.67%
	Multiple imputations	22	7.33%
	Imputation method not specified	11	3.67%
	MAR assumption stated	6	2.00%
	Weighted GEE	1	0.33%
	Maximum likelihood methods	2	0.67%
	Inverse probability weighing	3	1.00%
	Pattern mixture methods	2	0.67%
	Sensitivity analyses	2	0.67%
	Method not stated	219	73.00%
	*clustering standard errors (n=1), aggregation over clusters and subjects (n=1), Mantel Haenszel procedure over clusters (n=1), scale parameter for clusters (n=1), cluster-level stratification (n=1), Unstructured for clusters (n=1)		
Fixed or random effects for step (i.e., time)			
	Fixed	151	50.33%
	Random	4	1.33%
	Both	2	0.67%
	None	29	9.67%
	Not stated	114	38.00%

For fitting the step effect as fixed versus random, one hundred fifty-one (50.33%) trials used some form of fixed effect for step. Twenty-nine (9.67%) trials did not include a step effect in the model at all. Six (2.00%) trials specified the step effect as a random effect. One hundred fourteen (38.00%) trials, however, did not clearly indicate how the step effect was included in the model.

For choice of correlation structure, we organized our results by whether the investigators modeled correlation from both clusters and subjects, from clusters only, from subjects only, or when the level of correlation was unclear. For trials that adjusted for within-cluster and within-subject correlation, within-cluster correlation was handled through random effects while within-subject correlation was modeled as autoregressive (N=4, 1.33%), compound symmetric (N=1, 0.33%), random effects (N=30, 10.00%), random slopes (N=1, 0.33%), and unclear (N=1,0.33%). For trials that adjusted only for within-cluster correlation, correlation was accounted for with the following methods: bootstrapped standard errors (N=1,0.33%), clustering standard errors (N=1,0.33%), gamma frailty (N=1,0.33%), GEE (N=16,5.33%), random effects (N=102,34.00%), and robust standard errors (N=1,0.33%). For trials that only accounted for within-subject correlation, 2 (0.67%) trials used an autoregressive correlation structure and 9 (3.00%) trials used random effects. For the remaining trials, 30 (10.00%) trials treated observations as independent, 85 (28.33%) did not state how correlation was handled; 5 (1.67%) used random effects but without clearly specifying at which level; 4 (1.33%) were unclear about handling correlation; 1 (0.33%) used a Wilcoxon Signed Rank test; and 4 (1.33%) used other methods (see caption for Table 1).

For methods for missing data, two hundred nineteen trials (73.00%) did not clearly indicate how missing data were accounted for. Forty-two (14.00%) trials reported using some type of imputation. Twenty-one (7%) trials stated they were analyzing only complete cases. Two (0.67%) trials indicated use of maximum likelihood methods for missing data, but did not specify further. Four (1.33%) trials used inverse probability weighing. Two (0.67%) trials used pattern mixture models. Seven (2.33%) trials stated in the publication that the data were assumed to be either missing completely at random or missing at random. Two (0.67%) of trials stated that sensitivity analyses would be conducted for missing data, but were unclear about the exact approach.

From this systematic literature review, it is clear that detailed reporting on design and analysis of SWTs remains an issue with fifteen (5.00%) of trials not clearly specifying the overall design. Similarly, one hundred one (33.67%) trials failed to specify how temporal effects were controlled in their analysis. Issues such as heterogeneity of temporal effects or varying treatment effects might not exist in all trials. However, trials listed here as having treatment effect as *time-fixed* or *not stated* did not indicate whether any tests or sensitivity analyses were conducted concerning these issues. Out of the trials that specified the correlation structure, most trials used random effects or exchangeable correlation structure. Since linear mixed models are sensitive to specification of the correlation structure, more consideration should be given to choice of correlation structure. For missing data, LMMs provide unbiased estimates when the data are missing at random. This fact might explain why many trial authors (219 trials) did not feel the need to discuss missing data. Missing data are a more important concern when the analysis uses GEE, which is only unbiased when data are missing completely at random. Out of the forty-two (14%) studies that reported using imputation, only twenty-two (7.33%) clearly stated that multiple imputation was used. Seven trials used a fixed value (e.g. mean or median) imputation, and one trial used last value carry forward. Finally, one hundred fifty-one (50.33%) trials fit the step effect as some fixed effect, and one hundred fourteen (38%) did not clearly state how the step effect was fit. Twenty-none (9.67%) trials did not include a step effect.

2.7 APPLICATIONS

In this section, we present two real SWTs and discuss the parameters necessary for estimating the corresponding sample size and power via simulation.

2.7.1 Enhanced Recovery After Surgery (ERAS) Trial

Enhanced Recovery After Surgery (ERAS) protocols were developed and implemented around 2001 [15]. The protocols have been shown to improve postoperative recovery in colorectal surgery, combining newer anesthetic and minimally invasive surgery with evidence-based adjustments to facilitate re-validation. Enhanced recovery begins well before surgery through patient-doctor consultation with a focus on regular exercise, such as walking 30 minutes a day, improving eating habits and quitting smoking, in cases that apply. Other key elements may involve restricting the need for narcotics after surgery, allowing patients to drink clear liquids, including sports drinks, until two hours before procedures, limiting intravenous fluids during surgery, and encouraging patients to eat and walk around shortly after operations.

Let us suppose that a healthcare system plans to design a cross-sectional stepped wedge cluster randomized trial among patients undergo orthopedic surgeries. The goal is to compare patients length of stay (LOS) in hospital before and after implementing ERAS.

The trial will be conducted in 6 hospitals in the same healthcare system for 12 months. Patients undergo orthopedic surgeries who meet inclusion criteria during the enrollment period will be candidate participants. Each hospital will begin with usual recovery protocols (control phase) and transition to the ERAS (intervention phase) at a randomly assigned time (wedge). Each wedge involves a 3-month period. At each wedge, two hospitals will be randomly selected switching from control to intervention. From the historical data, patients with usual recovery protocols had an average length of stay 2.7 days. Between-site SD is estimated as 0.5 days.

To estimate the number of subjects who should be recruited for each hospital at each step, we performed sample size estimation via simulation. The outcome model is a Poisson model with a log link as shown in Equation (2.9).

$$log(\lambda_{ij}) = \mu + u_i \beta t_j + \theta trt_{ij} \tag{2.9}$$

such that

$$LOS_{k(ij)} \sim \text{Poisson}(\lambda_{ij})$$
$$u_i \sim N(0, \sigma_{cl}^2)$$

The model parameters are μ (the mean log LOS for subjects not on treatment, log(2.7) in this case), β (the temporal trend, 0 in this case), θ (the multiplicative effect of intervention on LOS, $log(2.2/2.7)$), and u_i are the random cluster effects with variance σ_{cl}^2. Based on a simulation with 1,000 Monte Carlo samples, we estimate that this trial should enroll 43 patients per cluster per step for a total of 1,032 patients throughout the entire study to achieve 80% power with a type I error rate of 0.05.

2.7.2 A Cirrhosis Trial

Cirrhosis is currently the 4th leading cause of death in the US among those ages 45-64 and the 6th leading cause of death among those ages 25-44 [3], and the mortality due to cirrhosis is increasing [14]. A fixed cohort stepped wedge cluster randomized trial is planned to evaluate a multilevel intervention to improve pain management among patients with cirrhosis.

The trial will include 5 clinical sites and recruit 50 patients for each site. All patients will be recruited at the study baseline and followed for 12 months. All sites will start with pre-intervention run-in. After the first month, the first randomly selected site will switch to a 2-month intervention, followed by the maintenance period. The second, third, fourth, and fifth randomly selected site will switch to the 2-month intervention at month 4, 6, 8, and 10, respectively. The maintenance period for the last randomized site will be 1 month. The procedure will continue until the end of the 5th site finishes the 2-month

intervention. Measurements for each patient will be taken at the end of each month (wedge) using the Postoperative Pain Self Management Behavior (PPSMB) scale. We utilized power simulations to estimate the minimal detectable difference between control and intervention. The outcome model is a mixed effect model as shown in Equation (2.10).

$$PPSMB_{k(i)j} = \mu + u_i + u_{k(i)} + \beta t_j + \theta trt_{ij} + e_{k(i)j} \tag{2.10}$$

such that

$$u_i \sim N(0, \sigma^2_{cl})$$
$$u_{k(i)} \sim N(0, \sigma^2_{cl})$$
$$e_{k(i)j} \sim N(0, \sigma^2)$$

The model parameters are μ (the mean PPSMB score for patients on control and baseline, 20.5 in this case), β (the temporal trend, generated as 0 in this case), and θ (the effect of intervention). The variance components are u_i (the random cluster intercepts with a variance of σ^2_{cl}), $u_{k(i)j}$ (the random subject intercepts with variance of σ^2_{subj}), and $e_{j(i)j}$ (the residuals with variance σ^2, $\sigma = 4$ in this case). Based upon simulations with 1,000 Monte Carlo samples, we estimate that the minimal detectable difference should be 0.60 assuming that the variance components σ_{cl} and σ_{subj} values might range from 0.10 to 0.20 to achieve 80% power with a 0.05 type one error rate.

2.8 DISCUSSION

SWT designs have a great variety of features in design and analysis. We proposed three terms for describing the overall design of a SWT: *fixed cohort*, *expanding cohort*, and *cross-sectional*. These three terms summarize the recruitment and exposure features of SWT designs. We discussed three design assumptions: homogeneity/heterogeneity of temporal effects, time-varying versus time-fixed treatment effect, and cluster-level versus subject-level temporal trends. For each of these design assumptions, we suggested appropriate statistical analysis models. We then discussed two different analytical assumptions: fitting of the step effect and choice of correlation structure.

Lastly, we provided the results of our systematic review of published SWT designs. The review describes the design type, design assumptions, and analytical assumptions for each trial. Based upon our systematic review of publications for SWTs, it is clear that many of these publications do not adequately discuss design and analytical aspects of their study. Without delineating such details, it is unclear whether these aspects were correctly considered when sample size and power estimations are presented. Most of these deficiencies in reporting features could be corrected by following the additional CONSORT guidelines for SWT designs as recommended by Hemming et al.[7]

We shall be releasing an R package for computing sample size and power for SWT designs via Monte Carlo simulations. The package will include *fixed cohort*, *expanding cohort*, and *cross-sectional* designs. In addition, the package will allow users to specify the design and analytical issues discussed in this paper including choice of correlation structure, whether the step effect is fixed or random, heterogeneity of temporal effects across clusters, time-varying treatment effects, and specification of cluster-level and subject-level temporal trends for expanding cohort designs.

3.0 DESIGN ASSUMPTIONS FOR STEPPED WEDGE CLUSTER RANDOMIZED TRIALS

ABSTRACT

The stepped wedge trial (SWT) is a type of cluster randomized trial with repeated measures and unidirectional crossover. The statistical literature for SWTs focuses on linear mixed models (LMMs), generalized linear mixed models (GLMMs), or generalized estimating equations (GEE) for analyzing SWTs. This paper focuses on the mean model specification for SWTs, particularly with respect to estimating and making inference on treatment effect. The primary interest is in understanding how these assumptions influence the estimation and inference for the effect of intervention. We conducted Monte Carlo simulation studies to study how the following three design assumptions influence estimation and inference for the intervention effect: (1) homogeneity or heterogeneity of temporal trends, (2) time-varying versus time-fixed intervention effects, and (3) cluster-level and subject-level temporal trends in expanding cohort SWT designs.

Since SWTs are often used for community-based and health policy interventions, they are particularly relevant to public health. Heterogeneity of temporal trends not accounted for in the model can result in inflated Type I error rate. Thus, recommendations based upon my simulation study can ensure that ineffective interventions are not wrongfully declared to be effective. Failure to account for time-varying treatment effect results in biased estimates for the effect of intervention. For that reason, checking for time-varying treatment effects is necessary to ensure the intervention effect is properly estimated. For expanding cohort designs, our simulation studies suggest that modeling the cluster-level temporal trend is important for proper estimation of the effect of intervention as well.

KEY WORDS : Cluster randomized trial, stepped wedge trial,temporal trend, time-varying treatment effect, .

3.1 INTRODUCTION

Stepped wedge trial (SWT) designs are a type of cluster randomized trial featuring repeated measures over time (also called steps) and unidirectional crossover from control to the intervention. These designs are often utilized for community-based interventions for public health initiatives or health policy. Justifications for using SWT designs include providing the intervention to all participating communities by the end of the trial, easier logistics for implementing the intervention when only a few cluster begin intervention at a given step, and the ability to formally control for temporal trends. While the primary interest is in estimation and inference for the effect of intervention, it is necessary to properly consider and model various design assumptions to do so accurately. First, investigators must identify which type of SWT design they are implementing. Previously, we described three main types of SWT designs: cross-sectional, fixed cohorts, and expanding cohorts.[cite paper 1] Cross-sectional SWT designs have continuous recruitment and subjects are only exposed to one step. Fixed cohort SWT designs start by recruiting all the subjects at baseline and then longitudinally follow them to the end of study or dropout. Expanding cohort SWT designs are similar to fixed cohorts, except that they allow subjects to continue to be recruited throughout the trial. In this paper, we focus on three main design assumptions to consider: heterogeneity/homogeneity of temporal trends, time-varying versus time-fixed treatment effect, and cluster-level and subject level temporal trends for expanding cohort SWTs. In Section 3.2, we provide the background information on the three design assumptions. In section 3.3, we present the results of Monte Carlo simulation studies we performed to assess the affect that misspecification of these design assumptions in the model has on the estimation and inference of the intervention effect.

3.2 DESIGN ASSUMPTIONS FOR STEPPED WEDGE TRIALS

3.2.1 Temporal effects are homogeneous or heterogeneous between clusters

Adjusting for background temporal effects need to be accounted for, which has been advised in the literature. When analyzing SWTs researchers often assumed that the background temporal effects are homogeneous across clusters. Hussey and Hughes[10] recommended fitting indicator variables for the step to account for temporal trends and noted that such temporal trends might be heterogeneous by cluster. We propose a statistical model that incorporates step effect heterogeneity by cluster as random coefficients for a parametric linear form of step. To study the effects that step effect heterogeneity might have on the estimation and inference of the fixed intervention effect, we performed Monte Carlo simulations studies in section 3.3.

3.2.2 Treatment effect is fixed or time-varying

Most analytical models for SWTs, including the influential paper by Hussey and Hughes [10], consider the effect of the intervention to be time-invariant. However, there are certain situations in which one might expect the effect of intervention to change over time. For example, the research staff might become more adept at delivering the intervention over time.

In Hussey and Hughes paper on the SWT design, they notion that the effect of the intervention might not be fully realized during a single step [10]. They suggested that in the middle of the study, we may use the design effect for intervention which is a factional value of treatment indicator (i.e., between 0 and 1) in place of the intervention design effect X_{ij} (with a value of 1) to reflect the incomplete achievement of full intervention effect. This topic was later revisited in another paper by Hughes et al. that provided a more detailed on the use of fractional X_{ij} values and the effects on power [9].

I investigated time-varying intervention effect for SWT designs using a statistical model with interaction terms between intervention effect and time. For simplicity, the time (i.e., step) effect was given a parametric linear form. Monte Carlo simulation studies were conducted to study estimation and inference for intervention effects when the intervention is time-varying. Data were simulated under various values of the main intervention effect parameter and the time by intervention interaction term parameter. Each set was analyzed both by the correct model and a model assuming the intervention effect is fixed over time.

3.2.3 Temporal Trends and Step Effects

It has been established in the stepped wedge literature that adjusting for background temporal trends is important for obtaining an unbiased estimate of the effect of intervention. However, expanding cohort designs have a unique issue for temporal effects compared to the other two types. In cross-sectional SWT designs without repeated measures, each subject is only affected by one step and has only one outcome measured. Thus for cross-sectional SWT designs without repeated measures, it is sufficient to adjust for any background secular trends by adjusting for the step at which the subject was exposed and measured. When repeated measures are taken on the same subject over the course of multiple steps, it seems likely that subject-level temporal trends (e.g., the natural history progression of a disease) might need to be accounted in the model to prevent confounding with the intervention effect.

For fixed cohort studies, Models (3.1) and (3.2) demonstrate that a model fitting indicator variables for the steps can control for both terms adequately. Consider in Model (3.1) the terms $\tau_{cl,l}$ are indicator variables for the cluster-level time trend while the terms $\tau_{subj,n}$ are indicator variables for the subject-level time trend.

Since each subject for a fixed cohort study is recruited at baseline, these time trends both occur on the same schedule, which can be seen by the shared index range from 2 to J. Thus, one can substitute the sum of these two indicators at every index value with another indicator β_l and still accurately account for both temporal trends.

$$E[y_{k(i)j}] = \mu + \theta trt_{ij} + \Sigma_{l=2}^{J} \tau_{cl,l} 1(j = l) + \Sigma_{n=2}^{J} \tau_{subj,n} 1(j = n) \tag{3.1}$$

Let $\tau_{cl,l} + \tau_{subj,l} = \beta_l$, then we get a model with both temporal trends accounted for by a single set of indicator variables.

$$E[y_{k(i)j}] = \mu + \theta trt_{ij} + \Sigma_{l=2}^{J} \beta_l 1(j = l) \tag{3.2}$$

On the other hand, consider the existence of cluster-level and subject-level time trends for an expanding cohort study. A model fitting both trends for an expanding cohort study is shown in Model (3.3). Note that t_{0k} represents the step during which subject k was recruited. The difference in index for the two temporal trends means that fitting indicator variables for the steps does not completely control for both temporal trends.

$$E[y_{k(i)j}] = \mu + \theta X_{ij} + \Sigma_{l=2}^{J} \tau_{cl,l} 1(j = l) + \Sigma_{n=2-t_{0k}+1}^{J} \tau_{subj,n} 1(j = n) \tag{3.3}$$

3.3 SIMULATION STUDIES

3.3.1 Cross-sectional SWT Designs

In this section, common assumptions regarding SWT designs will be discussed for open cross-sectional SWT designs. For each assumption, Monte Carlo simulation studies will be presented to study how violations of those assumptions affects the estimation and inference for the effect of intervention

3.3.1.1 Homogeneity/Heterogeneity of Temporal Trends Between Clusters

Monte Carlo simulations studies were performed with 1,000 samples to assess how heterogeneity of temporal trends across clusters can affect the estimation and inference for the effect of intervention. For simplicity, the temporal trend will be generated as linear. We chose to focus our simulation study on models with Gaussian-distributed error terms. Heterogeneity of the temporal effects across clusters were modeled through random slopes. The data generating model is provided in Equation 3.4.

$$y_{k(ij)} = (\beta_0 + b_{0i}) + \beta_1 trt_{ij} + (\beta_2 + b_{2i})t_j + e_{k(ij)} \tag{3.4}$$

for cluster i, time step j, and subject k where

$$\begin{bmatrix} b_{0i} \\ b_{2i} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2_{cl} & \sigma_{cl,step} \\ \sigma_{cl,step} & \sigma^2_{step} \end{bmatrix} \right)$$

$$e_{k(ij)} \sim N(0, \sigma^2)$$

Monte Carlo simulations were performed with the parameters outlined in Table 2.

For each combination of parameters, data will be generated according to Equation (3.4). Then, each simulated data set will be analyzed with three different models: the actual data generating model from Equation (3.4), a model that omits the random slope for each cluster shown in Equation (3.5), and a model that omits both the random slope and the random intercept for each cluster in Equation (3.6).

Table 2: Simulation parameters for the Monte Carlo study of heterogeneous temporal effects by cluster in a cross-sectional SWT design.

Parameter	Variable	Values
Total number of clusters	I	10
Total number of steps	J	$I+1$
Subjects per cluster per step	K	20
Model intercept	β_0	0.3
Mean time slope	β_2	0.25
Effect of intervention	β_1	$(0, 0.5, 1)$
Residual variance	σ^2	1.55^2
Random cluster effect variance	σ^2_{cl}	$(1.6, 2.3)$
Random slope variance	σ^2_{step}	$(1.3, 1.9)$
Random effect covariance	$\sigma_{cl,step}$	-1.8

$$y_{k(ij)} = (\beta_0 + b_{0i}) + \beta_1 trt_{ij} + \beta_2 t_j + e_{k(ij)} \tag{3.5}$$

for cluster i, time step j, and subject k where

$$b_{0i} \sim N(0, \sigma^2_{cl})$$
$$e_{k(i)j} \sim N(0, \sigma^2)$$

$$y_{k(ij)} = \beta_0 + \beta_1 trt_{ij} + \beta_2 t_j + e_{k(ij)} \tag{3.6}$$

for cluster i, time step j, and subject k where

$$e_{k(i)j} \sim N(0, \sigma^2)$$

For each analysis, the the bias, average model-based standard error, Monte Carlo standard error, coverage probability, and the power/Type I error rate are reported in Tables (3) - (5). The estimation of the treatment effect parameter β_1 was not bias for either the model without a random slope or the model without a random intercept for each cluster. Estimation of

Table 3: Monte Carlo simulation for heterogeneity of temporal effects by cluster in a cross-sectional SWT design results when analyzing with the correct model (random slopes and intercepts).

				True model				
Theta	σ_cl	σ_step	σ_cl,step	Bias	Model SE	MC SE	Coverage	Power / Type I error
0	1.6	1.3	-1.8	0.0055	0.1220	0.1218	0.9480	0.0520
0.5	1.6	1.3	-1.8	0.0016	0.1220	0.1213	0.9460	0.9820
1	1.6	1.3	-1.8	-0.0019	0.1219	0.1260	0.9410	1.0000
0	2.3	1.3	-1.8	-0.0011	0.1236	0.1229	0.9550	0.0450
0.5	2.3	1.3	-1.8	-0.0052	0.1236	0.1238	0.9490	0.9810
1	2.3	1.3	-1.8	0.0037	0.1235	0.1273	0.9460	1.0000
0	1.6	1.9	-1.8	0.0060	0.1232	0.1273	0.9360	0.0640
0.5	1.6	1.9	-1.8	0.0015	0.1231	0.1226	0.9510	0.9840
1	1.6	1.9	-1.8	-0.0032	0.1231	0.1247	0.9390	1.0000
0	2.3	1.9	-1.8	-0.0036	0.1237	0.1211	0.9580	0.0420
0.5	2.3	1.9	-1.8	-0.0019	0.1237	0.1249	0.9480	0.9820
1	2.3	1.9	-1.8	0.0010	0.1238	0.1235	0.9480	1.0000

Table 4: Monte Carlo simulation for heterogeneity of temporal effects by cluster in a cross-sectional SWT design results when analyzing with the the model without random slopes.

				No Random Slope				
Theta	σ_cl	σ_step	σ_cl,step	Bias	Model SE	MC SE	Coverage	Power / Type I error
0	1.6	1.3	-1.8	0.0214	0.3048	1.6839	0.2680	0.7320
0.5	1.6	1.3	-1.8	-0.0293	0.3073	1.6424	0.2720	0.7200
1	1.6	1.3	-1.8	0.0072	0.3064	1.6569	0.2790	0.7760
0	2.3	1.3	-1.8	-0.0032	0.3023	1.6351	0.2783	0.7217
0.5	2.3	1.3	-1.8	-0.0437	0.3058	1.6665	0.2780	0.7170
1	2.3	1.3	-1.8	-0.0190	0.3054	1.6077	0.2520	0.7520
0	1.6	1.9	-1.8	0.1055	0.4316	2.5148	0.2390	0.7600
0.5	1.6	1.9	-1.8	0.1086	0.4267	2.3152	0.2840	0.7380
1	1.6	1.9	-1.8	0.1449	0.4236	2.3883	0.2400	0.7820
0	2.3	1.9	-1.8	0.0310	0.4265	2.4321	0.2610	0.7390
0.5	2.3	1.9	-1.8	-0.0329	0.4262	2.3530	0.2700	0.7040
1	2.3	1.9	-1.8	-0.0028	0.4236	2.4738	0.2510	0.7770

Table 5: Monte Carlo simulation for heterogeneity of temporal effects by cluster in a cross-sectional SWT design results when analyzing with the the model with neither random slopes nor random intercepts.

				No Random Intercept				
Theta	σ_cl	σ_step	σ_cl,step	Bias	Model SE	MC SE	Coverage	Power / Type I error
0	1.6	1.3	-1.8	-0.1423	0.3914	3.6310	0.1690	0.8310
0.5	1.6	1.3	-1.8	0.0651	0.3954	3.5761	0.1730	0.8330
1	1.6	1.3	-1.8	-0.0759	0.3941	3.7203	0.1630	0.8380
0	2.3	1.3	-1.8	0.2489	0.3970	3.5993	0.1600	0.8400
0.5	2.3	1.3	-1.8	0.0022	0.4016	3.7066	0.1670	0.8420
1	2.3	1.3	-1.8	-0.0173	0.4020	3.7284	0.1430	0.8460
0	1.6	1.9	-1.8	0.3041	0.6147	6.0466	0.1390	0.8610
0.5	1.6	1.9	-1.8	0.1285	0.6088	5.8496	0.1490	0.8590
1	1.6	1.9	-1.8	-0.0499	0.6030	5.8908	0.1650	0.8300
0	2.3	1.9	-1.8	0.0649	0.6127	6.2648	0.1360	0.8640
0.5	2.3	1.9	-1.8	0.0949	0.6148	5.8950	0.1550	0.8390
1	2.3	1.9	-1.8	0.2072	0.6105	5.7845	0.1570	0.8450

the standard error for that parameter, however, was deflated compared to the Monte Carlo standard error for the models without random slopes or without random intercepts whenever the random effect variances were nonzero. Thus, the misspecified models also resulted in smaller coverage probability for the treatment effect. When the treatment effect was set to zero, the power column represents Type I error rate. The misspecified models both exhibit increased Type I error rate for testing the treatment effect.

3.3.1.2 Fixed Versus Time-varying Treatment Effects The following model will be used to generate data for a cross-sectional SWT design with a time-varying effect of intervention with Gaussian-distributed error terms.

$$y_{k(ij)} = (\beta_0 + b_{0i}) + \beta_1 trt_{ij} + \beta_2 t_j + \beta_3 trt_{ij} t_j + e_{k(ij)} \tag{3.7}$$

for cluster i, time step j, and subject k where

$$b_{0i} \sim N(0, \sigma^2_{cl})$$
$$e_{k(ij)} \sim N(0, \sigma^2)$$

Table 6: Simulation parameters for the Monte Carlo study of time-varying treatment effect in a cross-sectional SWT design.

Parameter	Variable	Values
Total number of clusters	I	10
Total number of steps	J	$I+1$
Subjects per cluster per step	K	20
Model intercept	β_0	0.3
Effect of intervention	β_1	$(-1, 0, 0.5, 1)$
Mean time slope	β_2	0.25
Intervention by Time Interaction	β_3	$(-0.10, -0.05, 0, 0.05, 0.10)$
Residual variance	σ^2	1.55^2
Random cluster effect variance	σ^2_{cl}	0.0777^2

Monte Carlo simulations were conducted using 1,000 samples according to the parameters outlined in Table 6.

Each simulated data set was then analyzed by two different models: the correct data generating model from Equation (3.7) and a model that omits the intervention by time interaction term as shown in Equation (3.8).

$$y_{k(ij)} = (\beta_0 + b_{0i}) + \beta_1 trt_{ij} + \beta_2 t_j + e_{k(ij)} \tag{3.8}$$

for cluster i, time step j, and subject k where

$$b_{0i} \sim N(0, \sigma^2_{cl})$$
$$e_{k(ij)} \sim N(0, \sigma^2)$$

For each choice of parameters, the mean bias, mean model-based standard error, Monte Carlo standard error, coverage probability, and power/Type I error rate for the effect of intervention are reported. For the time-varying model, the power/Type I error rate will be reported for an early tstep (i.e., the second step), a middle step ($0.5(J-2)+2$ to the

nearest step), and the last step. These results are presented in Tables (7) and (8). The power for the contrast tests exhibit a monotone relationship over time when β_1 and β_3 match in sign. When β_1 and β_3 differ in sign, then . When the signs of β_1 and β_3 differ, the contrast tests show a non-monotonic pattern over time. The time-varying model results show no bias issues. However, the Monte Carlo standard error suggests that the model-based standard error is deflated. The time-fixed treatment effect models do exhibit issues with bias, the sign of which clearly matching the sign of the omitted β_3 interaction term. Monte Carlo standard error suggests that the estimation of the treatment parameter estimate standard error was unbiased. Coverage probability decreases in alignment with the bias issues. Power is decreased especially for the scenarios in which the signs of β_1 and β_3 differ.

3.3.2 Fixed Cohort SWT Designs

In this section, common assumptions regarding SWT designs will be discussed for fixed cohort SWT designs. For each assumption, Monte Carlo simulation studies will be presented to study how violations of those assumptions affects the estimation and inference for the effect of intervention

3.3.2.1 Homogeneity/Heterogeneity of Temporal Trends Between Clusters Monte Carlo simulations studies were performed to assess how heterogeneity of temporal trends across clusters can affect the estimation and inference for the effect of intervention. For simplicity, the temporal trend will be generated as linear. For our simulations, we chose to study models with Gaussian-distributed error terms. Heterogeneity of the temporal effects across clusters were modeled through random slopes. The data generating model is provided in Equation 3.9.

Table 7: Monte Carlo simulation results for time-varying treatment effects in a cross-sectional SWT design when analyzing with a time-varying treatment effect model.

						Time-varying		
Beta_1	**Beta_3**	**Bias**	**Model SE**	**MC SE**	**Coverage**	**Test Early**	**Test Mid**	**Test Late**
-1	-0.1	-0.0022	0.0935	0.1905	0.9520	1.0000	1.0000	1.0000
-0.5	-0.1	-0.0045	0.0933	0.1944	0.9340	0.9980	1.0000	1.0000
0	-0.1	0.0039	0.0937	0.1985	0.9340	0.2950	1.0000	1.0000
0.5	-0.1	0.0042	0.0935	0.1947	0.9460	0.5660	0.1740	0.9440
1	-0.1	0.0013	0.0937	0.1955	0.9470	1.0000	0.9900	0.0940
-1	-0.05	0.0023	0.0937	0.1912	0.9410	1.0000	1.0000	1.0000
-0.5	-0.05	0.0036	0.0935	0.1922	0.9479	0.9830	1.0000	1.0000
0	-0.05	-0.0045	0.0935	0.1938	0.9459	0.1231	0.8969	0.9199
0.5	-0.05	-0.0027	0.0936	0.1897	0.9459	0.7886	0.5511	0.0631
1	-0.05	0.0111	0.0936	0.1824	0.9580	1.0000	1.0000	0.7680
-1	0	-0.0080	0.0935	0.1895	0.9460	1.0000	1.0000	1.0000
-0.5	0	0.0029	0.0934	0.1903	0.9389	0.9269	1.0000	0.8457
0	0	0.0011	0.0935	0.1895	0.9449	0.0501	0.0571	0.0541
0.5	0	-0.0040	0.0938	0.1894	0.9510	0.9320	0.9990	0.8680
1	0	0.0031	0.0935	0.1928	0.9420	1.0000	1.0000	1.0000
-1	0.05	-0.0039	0.0933	0.1916	0.9450	1.0000	1.0000	0.7850
-0.5	0.05	-0.0027	0.0936	0.1942	0.9419	0.8016	0.6012	0.0701
0	0.05	-0.0008	0.0934	0.1940	0.9389	0.1191	0.9049	0.9159
0.5	0.05	0.0015	0.0936	0.1795	0.9650	0.9960	1.0000	1.0000
1	0.05	0.0059	0.0934	0.1914	0.9450	1.0000	1.0000	1.0000
-1	0.1	0.0001	0.0938	0.1939	0.9480	1.0000	0.9900	0.0910
-0.5	0.1	0.0020	0.0937	0.1902	0.9489	0.5566	0.1942	0.9469
0	0.1	-0.0022	0.0935	0.1934	0.9570	0.2980	1.0000	1.0000
0.5	0.1	0.0030	0.0935	0.1925	0.9490	0.9990	1.0000	1.0000
1	0.1	0.0028	0.0936	0.1948	0.9260	1.0000	1.0000	1.0000

Table 8: Monte Carlo simulation results for time-varying treatment effects in a cross-sectional SWT design when analyzing with a time-fixed treatment effect model.

			Time-fixed			
Beta_1	**Beta_3**	**Bias**	**Model SE**	**MC SE**	**Coverage**	**Power**
-1	-0.1	-0.5971	0.0931	0.0961	0.0000	1.0000
-0.5	-0.1	-0.5932	0.0928	0.0995	0.0000	1.0000
0	-0.1	-0.6004	0.0935	0.0996	0.0000	1.0000
0.5	-0.1	-0.5971	0.0932	0.0921	0.0000	0.1660
1	-0.1	-0.5966	0.0936	0.0943	0.0000	0.9890
-1	-0.05	-0.2965	0.0926	0.0971	0.1280	1.0000
-0.5	-0.05	-0.2988	0.0924	0.0954	0.1160	1.0000
0	-0.05	-0.3030	0.0923	0.0951	0.1041	0.8899
0.5	-0.05	-0.3038	0.0926	0.0940	0.1050	0.5440
1	-0.05	-0.2972	0.0927	0.0942	0.1130	1.0000
-1	0	0.0002	0.0922	0.0922	0.9459	1.0000
-0.5	0	0.0017	0.0920	0.0975	0.9239	1.0000
0	0	0.0001	0.0920	0.0919	0.9439	0.0561
0.5	0	-0.0024	0.0925	0.0936	0.9470	0.9990
1	0	-0.0018	0.0923	0.0947	0.9389	1.0000
-1	0.05	0.2971	0.0920	0.0925	0.1070	1.0000
-0.5	0.05	0.2967	0.0924	0.0961	0.1130	0.5870
0	0.05	0.3004	0.0923	0.0927	0.0970	0.9020
0.5	0.05	0.3055	0.0925	0.0920	0.0940	1.0000
1	0.05	0.3002	0.0921	0.0935	0.1072	1.0000
-1	0.1	0.5981	0.0934	0.0945	0.0000	0.9900
-0.5	0.1	0.5985	0.0932	0.0967	0.0000	0.1892
0	0.1	0.6028	0.0932	0.0946	0.0000	1.0000
0.5	0.1	0.5988	0.0932	0.0954	0.0000	1.0000
1	0.1	0.6023	0.0932	0.0947	0.0000	1.0000

Table 9: Simulation parameters for the Monte Carlo study of heterogeneous temporal effects by cluster in a fixed cohort SWT design.

Parameter	Variable	Values
Total number of clusters	I	10
Total number of steps	J	$I+1$
Subjects per cluster per step	K	20
Model intercept	β_0	0.3
Mean time slope	β_2	0.25
Effect of intervention	β_1	$(0, 0.5, 1)$
Residual variance	σ^2	1.55^2
Random cluster effect variance	σ^2_{cl}	$(1.6, 2.3)$
Random slope variance	σ^2_{step}	$(1.3, 1.9)$
Random effect covariance	$\sigma_{cl,step}$	-1.8
Random subject effect variance	σ^2_{subj}	0.8

$$y_{k(i)j} = (\beta_0 + b_{0i} + b_{0k}) + \beta_1 trt_{ij} + (\beta_2 + b_{2i})t_j + e_{k(i)j} \tag{3.9}$$

for cluster i, time step j, and subject k where

$$\begin{bmatrix} b_{0i} \\ b_{2i} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2_{cl} & \sigma_{cl,step} \\ \sigma_{cl,step} & \sigma^2_{step} \end{bmatrix} \right)$$

$$e_{k(i)j} \sim N(0, \sigma^2)$$

Monte Carlo simulations were conducted with 1,000 samples using the parameters outlined in Table 9.

For each combination of parameters, data will be generated according to Equation (3.9). Then, each simulated data set will be analyzed with three different models: the actual data generating model from Equation (3.9), a model that omits the random slope for each cluster shown in Equation (3.10), and a model that omits both the random slope and the random intercept for each cluster in Equation (3.11).

$$y_{k(i)j} = (\beta_0 + b_{0i} + b_{0k}) + \beta_1 trt_{ij} + \beta_2 t_j + e_{k(i)j} \tag{3.10}$$

for cluster i, time step j, and subject k where

$$\begin{aligned} b_{0i} &\sim N(0, \sigma^2_{cl}) \\ b_{0k} &\sim N(0, \sigma^2_{subj}) \\ e_{k(i)j} &\sim N(0, \sigma^2) \end{aligned}$$

$$y_{k(i)j} = (\beta_0 + b_{0k}) + \beta_1 trt_{ij} + \beta_2 t_j + e_{k(i)j} \tag{3.11}$$

for cluster i, time step j, and subject k where

$$\begin{aligned} b_{0k} &\sim N(0, \sigma^2_{subj}) \\ e_{k(i)j} &\sim N(0, \sigma^2) \end{aligned}$$

For each analysis, the the bias, average model-based standard error, Monte Carlo standard error, coverage probability, and the power/Type I error rate are reported in Tables (10) - (12).The estimation of the treatment effect parameter β_1 was not bias for either the model without a random slope or the model without a random intercept for each cluster. Estimation of the standard error for that parameter, however, was deflated compared to the Monte Carlo standard error for the models without random slopes or without random intercepts whenever the random effect variances were nonzero. Thus, the misspecified models also resulted in smaller coverage probability for the treatment effect. When the treatment effect was set to zero, the power column represents Type I error rate. The misspecified models both exhibit increased Type I error rate for testing the treatment effect.

Table 10: Monte Carlo simulation for heterogeneity of temporal effects by cluster in a fixed cohort SWT design results when analyzing with the correct model (random slopes and intercepts).

				True model				
Theta	σ_cl	σ_step	σ_cl,step	Bias	Model SE	MC SE	Coverage	Power / Type I error
0	1.6	1.3	-1.8	0.0053	0.1223	0.1251	0.9450	0.0550
0.5	1.6	1.3	-1.8	-0.0010	0.1222	0.1266	0.9380	0.9780
1	1.6	1.3	-1.8	0.0003	0.1221	0.1248	0.9420	1.0000
0	2.3	1.3	-1.8	0.0023	0.1236	0.1244	0.9520	0.0480
0.5	2.3	1.3	-1.8	0.0024	0.1237	0.1224	0.9500	0.9790
1	2.3	1.3	-1.8	0.0027	0.1237	0.1222	0.9590	1.0000
0	1.6	1.9	-1.8	-0.0015	0.1233	0.1230	0.9500	0.0500
0.5	1.6	1.9	-1.8	0.0065	0.1234	0.1255	0.9470	0.9840
1	1.6	1.9	-1.8	0.0004	0.1233	0.1239	0.9530	1.0000
0	2.3	1.9	-1.8	0.0032	0.1238	0.1225	0.9570	0.0430
0.5	2.3	1.9	-1.8	0.0100	0.1238	0.1210	0.9520	0.9800
1	2.3	1.9	-1.8	0.0031	0.1237	0.1235	0.9530	1.0000

Table 11: Monte Carlo simulation for heterogeneity of temporal effects by cluster in a fixed cohort SWT design results when analyzing with the the model without random slopes.

				No Random Slope				
Theta	σ_cl	σ_step	σ_cl,step	Bias	Model SE	MC SE	Coverage	Power / Type I error
0	1.6	1.3	-1.8	-0.0125	0.3082	1.6620	0.2793	0.7207
0.5	1.6	1.3	-1.8	0.1338	0.3120	1.6986	0.2840	0.7490
1	1.6	1.3	-1.8	-0.0067	0.3120	1.6567	0.2740	0.7590
0	2.3	1.3	-1.8	-0.0314	0.3095	1.6877	0.2833	0.7167
0.5	2.3	1.3	-1.8	0.0030	0.3096	1.6667	0.2583	0.7457
1	2.3	1.3	-1.8	-0.0166	0.3082	1.6713	0.2833	0.7548
0	1.6	1.9	-1.8	-0.0110	0.4323	2.4450	0.2633	0.7367
0.5	1.6	1.9	-1.8	-0.0611	0.4323	2.4528	0.2530	0.7460
1	1.6	1.9	-1.8	-0.1075	0.4309	2.4527	0.2653	0.7608
0	2.3	1.9	-1.8	0.0444	0.4257	2.4100	0.2750	0.7250
0.5	2.3	1.9	-1.8	-0.0683	0.4302	2.4924	0.2420	0.7650
1	2.3	1.9	-1.8	0.0825	0.4284	2.3610	0.2760	0.7660

Table 12: Monte Carlo simulation for heterogeneity of temporal effects by cluster in a fixed cohort SWT design results when analyzing with the the model with neither random slopes nor random intercepts.

				No Random Intercept				
Theta	σ_cl	σ_step	σ_cl,step	Bias	Model SE	MC SE	Coverage	Power / Type I error
0	1.6	1.3	-1.8	-0.0130	0.3092	1.6575	0.2840	0.7160
0.5	1.6	1.3	-1.8	0.1282	0.3131	1.6821	0.2630	0.7430
1	1.6	1.3	-1.8	-0.0113	0.3131	1.6381	0.2660	0.7520
0	2.3	1.3	-1.8	-0.0376	0.3106	1.6775	0.2740	0.7250
0.5	2.3	1.3	-1.8	0.0028	0.3108	1.6529	0.2790	0.7530
1	2.3	1.3	-1.8	-0.0171	0.3093	1.6382	0.3010	0.7630
0	1.6	1.9	-1.8	-0.0002	0.4405	2.4163	0.2660	0.7340
0.5	1.6	1.9	-1.8	-0.0635	0.4406	2.4199	0.2700	0.7320
1	1.6	1.9	-1.8	-0.0879	0.4392	2.4299	0.2810	0.7540
0	2.3	1.9	-1.8	0.0491	0.4338	2.3858	0.2640	0.7360
0.5	2.3	1.9	-1.8	-0.0551	0.4385	2.4698	0.2450	0.7530
1	2.3	1.9	-1.8	0.0748	0.4366	2.3293	0.2900	0.7650

3.3.2.2 Fixed Versus Time-varying Treatment Effects The following model will be used to generate data for a fixed cohort SWT design with a time-varying effect of intervention with Gaussian-distributed error terms.

$$y_{k(i)j} = (\beta_0 + b_{0i} + b_{0k(i)}) + \beta_1 trt_{ij} + \beta_2 t_j + \beta_3 trt_{ij} t_j + e_{k(i)j} \tag{3.12}$$

for cluster i, time step j, and subject k where

$$b_{0i} \sim N(0, \sigma^2_{cl})$$
$$b_{0k} \sim N(0, \sigma^2_{subj})$$
$$e_{k(ij)} \sim N(0, \sigma^2)$$

Table 13: Simulation parameters for the Monte Carlo study of time-varying treatment effect in a fixed cohort SWT design.

Parameter	Variable	Values
Total number of clusters	I	10
Total number of steps	J	$I+1$
Subjects per cluster per step	K	20
Model intercept	β_0	0.3
Effect of intervention	β_1	$(-1, 0, 0.5, 1)$
Mean time slope	β_2	0.25
Intervention by Time Interaction	β_3	$(-0.10, -0.05, 0, 0.05, 0.10)$
Residual variance	σ^2	1.55^2
Random cluster effect variance	σ^2_{cl}	0.0777^2
Random subject effect variance	σ^2_{subj}	0.1

Monte Carlo simulations were performed with 1,000 samples using the parameters outlined in Table 13. Each simulated data set was then analyzed by two different models: the correct data generating model from Equation (3.12) and a model that omits the intervention by time interaction term as shown in Equation (3.13).

$$y_{k(ij)} = (\beta_0 + b_{0i}) + \beta_1 trt_{ij} + \beta_2 t_j + e_{k(ij)} \tag{3.13}$$

for cluster i, time step j, and subject k where

$$b_{0i} \sim N(0, \sigma^2_{cl})$$
$$e_{k(ij)} \sim N(0, \sigma^2)$$

For each choice of parameters, the mean bias, mean model-based standard error, Monte Carlo standard error, coverage probability, and power/Type I error rate for the effect of intervention are reported. For the time-varying model, the power/Type I error rate will be reported for an early tstep (i.e., the second step), a middle step ($0.5(J-2)+2$ to the nearest step), and the last step. These results are presented in Tables (14) and (15). For

the time-varying model, the power/Type I error rate will be reported for an early tstep (i.e., the second step), a middle step ($0.5(J-2)+2$ to the nearest step), and the last step. These results are presented in Tables (7) and (8). The power for the contrast tests exhibit a monotone relationship over time when β_1 and $\beta)3$ match in sign. When β_1 and β_3 differ in sign, then . When the signs of β_1 and β_3 differ, the contrast tests show a non-monotonic pattern over time. The time-varying model results show no bias issues. However, the Monte Carlo standard error suggests that the model-based standard error is deflated. The time-fixed treatment effect models do exhibit issues with bias, the sign of which clearly matching the sign of the omitted β_3 interaction term. Monte Carlo standard error suggests that the estimation of the treatment parameter estimate standard error was unbiased. Coverage probability decreases in alignment with the bias issues. Power is decreased especially for the scenarios in which the signs of β_1 and β_3 differ.

3.3.3 Expanding cohort SWT Designs

3.3.3.1 Homogeneity/Heterogeneity of Temporal Trends Between Clusters Monte Carlo simulations studies were performed to assess how heterogeneity of temporal trends across clusters can affect the estimation and inference for the effect of intervention. For simplicity, the temporal trend will be generated as linear. We chose to study models with Gaussian-distibuted error terms. Heterogeneity of the temporal effects across clusters were modeled through random slopes. The data generating model is provided in Equation 3.14.

$$y_{k(i)j} = (\beta_0 + b_{0i} + b_{0k}) + \beta_1 trt_{ij} + (\beta_2 + b_{2i})t_j + e_{k(i)j} \tag{3.14}$$

for cluster i, time step j, and subject k where

$$\begin{bmatrix} b_{0i} \\ b_{2i} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2_{cl} & \sigma_{cl,step} \\ \sigma_{cl,step} & \sigma^2_{step} \end{bmatrix} \right)$$

$$e_{k(i)j} \sim N(0, \sigma^2)$$

Table 14: Monte Carlo simulation results for time-varying treatment effects in a fixed cohort SWT design when analyzing with a time-varying treatment effect model.

						Time-varying		
Beta_1	**Beta_3**	**Bias**	**Model SE**	**MC SE**	**Coverage**	**Test Early**	**Test Mid**	**Test Late**
-1	-0.1	-0.0007	0.0938	0.1835	0.9600	1.0000	1.0000	1.0000
-0.5	-0.1	-0.0044	0.0938	0.1872	0.9490	0.9980	1.0000	1.0000
0	-0.1	-0.0034	0.0940	0.1913	0.9409	0.2863	1.0000	1.0000
0.5	-0.1	0.0127	0.0939	0.1905	0.9499	0.5596	0.1812	0.9560
1	-0.1	-0.0035	0.0937	0.1859	0.9469	1.0000	0.9850	0.1001
-1	-0.05	-0.0014	0.0939	0.1892	0.9560	1.0000	1.0000	1.0000
-0.5	-0.05	0.0061	0.0937	0.1856	0.9520	0.9880	1.0000	1.0000
0	-0.05	0.0013	0.0937	0.1893	0.9410	0.1080	0.8890	0.9080
0.5	-0.05	-0.0015	0.0941	0.1860	0.9530	0.7960	0.5550	0.0680
1	-0.05	-0.0053	0.0936	0.1923	0.9540	1.0000	1.0000	0.7528
-1	0	-0.0001	0.0939	0.1895	0.9469	1.0000	1.0000	1.0000
-0.5	0	-0.0012	0.0938	0.1926	0.9469	0.9309	1.0000	0.8478
0	0	0.0010	0.0940	0.1843	0.9520	0.0570	0.0550	0.0610
0.5	0	0.0011	0.0939	0.1914	0.9420	0.9350	0.9990	0.8450
1	0	0.0041	0.0937	0.1916	0.9399	1.0000	1.0000	1.0000
-1	0.05	0.0052	0.0937	0.2003	0.9340	1.0000	1.0000	0.7800
-0.5	0.05	0.0011	0.0938	0.1872	0.9570	0.7930	0.5740	0.0780
0	0.05	0.0043	0.0939	0.1835	0.9580	0.0990	0.8830	0.9110
0.5	0.05	0.0042	0.0938	0.1853	0.9480	0.9880	1.0000	1.0000
1	0.05	0.0103	0.0934	0.1951	0.9360	1.0000	1.0000	1.0000
-1	0.1	-0.0003	0.0939	0.1933	0.9360	1.0000	0.9910	0.0970
-0.5	0.1	-0.0090	0.0938	0.1926	0.9450	0.5670	0.1960	0.9590
0	0.1	0.0139	0.0938	0.1887	0.9510	0.3210	1.0000	1.0000
0.5	0.1	0.0014	0.0938	0.1958	0.9360	0.9990	1.0000	1.0000
1	0.1	0.0083	0.0940	0.1944	0.9460	1.0000	1.0000	1.0000

Table 15: Monte Carlo simulation results for time-varying treatment effects in a fixed cohort SWT design when analyzing with a time-fixed treatment effect model.

				Time-fixed		
Beta_1	**Beta_3**	**Bias**	**Model SE**	**MC SE**	**Coverage**	**Power**
-1	-0.1	-0.5978	0.0934	0.0917	0.0000	1.0000
-0.5	-0.1	-0.5989	0.0936	0.0937	0.0000	1.0000
0	-0.1	-0.6012	0.0940	0.0944	0.0000	1.0000
0.5	-0.1	-0.5982	0.0938	0.0913	0.0000	0.1702
1	-0.1	-0.6030	0.0935	0.0961	0.0000	0.9830
-1	-0.05	-0.3043	0.0929	0.0957	0.0990	1.0000
-0.5	-0.05	-0.2988	0.0928	0.0919	0.0981	1.0000
0	-0.05	-0.2976	0.0929	0.0947	0.1101	0.8829
0.5	-0.05	-0.3011	0.0931	0.0954	0.1210	0.5470
1	-0.05	-0.3089	0.0927	0.0953	0.0950	1.0000
-1	0	-0.0060	0.0927	0.0975	0.9320	1.0000
-0.5	0	0.0043	0.0928	0.0942	0.9430	1.0000
0	0	-0.0043	0.0929	0.0973	0.9460	0.0530
0.5	0	-0.0011	0.0927	0.0992	0.9310	0.9990
1	0	-0.0004	0.0926	0.0971	0.9330	1.0000
-1	0.05	0.3009	0.0930	0.0953	0.1021	1.0000
-0.5	0.05	0.3011	0.0929	0.0963	0.1130	0.5620
0	0.05	0.2995	0.0931	0.0960	0.1200	0.8740
0.5	0.05	0.3003	0.0928	0.0985	0.1310	1.0000
1	0.05	0.3033	0.0925	0.0954	0.0960	1.0000
-1	0.1	0.5987	0.0940	0.0960	0.0000	0.9910
-0.5	0.1	0.5991	0.0940	0.0924	0.0000	0.1860
0	0.1	0.6031	0.0936	0.0936	0.0000	1.0000
0.5	0.1	0.5985	0.0937	0.0960	0.0000	1.0000
1	0.1	0.6006	0.0938	0.0931	0.0000	1.0000

Table 16: Simulation parameters for the Monte Carlo study of heterogeneous temporal effects by cluster in an expanding cohort SWT design.

Parameter	Variable	Values
Total number of clusters	I	10
Total number of steps	J	$I+1$
Subjects per cluster per step	K	20
Model intercept	β_0	0.3
Mean time slope	β_2	0.25
Effect of intervention	β_1	$(0, 0.5, 1)$
Residual variance	σ^2	1.55^2
Random cluster effect variance	σ^2_{cl}	$(1.6, 2.3)$
Random slope variance	σ^2_{step}	$(1.3, 1.9)$
Random effect covariance	$\sigma_{cl,step}$	-1.8
Random subject effect variance	σ^2_{subj}	0.8

To simulate enrollment of subjects after the first step, a multinomial random variable was generated for each subject with probability of enrolling during the first step equaling 0.6 and the probability of enrolling during all later steps having the uniform probability equal to $\frac{0.4}{J-1}$.

Monte Carlo simulations were performed with 1,000 samples using the parameters outlined in Table 16. For each combination of parameters, data will be generated according to Equation (3.14). Then, each simulated data set will be analyzed with three different models: the actual data generating model from Equation (3.14), a model that omits the random slope for each cluster shown in Equation (3.15), and a model that omits both the random slope and the random intercept for each cluster in Equation (3.16).

$$y_{k(i)j} = (\beta_0 + b_{0i} + b_{0k}) + \beta_1 trt_{ij} + \beta_2 t_j + e_{k(i)j} \tag{3.15}$$

for cluster i, time step j, and subject k where

$$b_{0i} \sim N(0, \sigma^2_{cl})$$
$$b_{0k} \sim N(0, \sigma^2_{subj})$$
$$e_{k(i)j} \sim N(0, \sigma^2)$$

$$y_{k(i)j} = (\beta_0 + b_{0k}) + \beta_1 trt_{ij} + \beta_2 t_j + e_{k(i)j} \tag{3.16}$$

for cluster i, time step j, and subject k where

$$b_{0k} \sim N(0, \sigma^2_{subj})$$
$$e_{k(i)j} \sim N(0, \sigma^2)$$

For each analysis, the the bias, average model-based standard error, Monte Carlo standard error, coverage probability, and the power/Type I error rate are reported in Tables (17) - (19). The estimation of the treatment effect parameter β_1 was not bias for either the model without a random slope or the model without a random intercept for each cluster. Estimation of the standard error for that parameter, however, was deflated compared to the Monte Carlo standard error for the models without random slopes or without random intercepts whenever the random effect variances were nonzero. Thus, the misspecified models also resulted in smaller coverage probability for the treatment effect. When the treatment effect was set to zero, the power column represents Type I error rate. The misspecified models both exhibit increased Type I error rate for testing the treatment effect.

3.3.3.2 Fixed Versus Time-varying Treatment Effects The following model will be used to generate data for an expanding cohort SWT design with a time-varying effect of intervention with Gaussian-distributed error terms.

Table 17: Monte Carlo simulation for heterogeneity of temporal effects by cluster in an expanding cohort SWT design results when analyzing with the correct model (random slopes and intercepts).

				True model				
Theta	σ_cl	σ_step	σ_cl,step	Bias	Model SE	MC SE	Coverage	Power / Type I error
0	1.6	1.3	-1.8	0.0257	0.3338	0.3409	0.9354	0.0635
0.5	1.6	1.3	-1.8	-0.0142	0.3335	0.3315	0.9506	0.2986
1	1.6	1.3	-1.8	0.0094	0.3342	0.3319	0.9466	0.8632
0	2.3	1.3	-1.8	0.0070	0.3355	0.3443	0.9424	0.0555
0.5	2.3	1.3	-1.8	-0.0337	0.3338	0.3398	0.9435	0.3068
1	2.3	1.3	-1.8	0.0106	0.3347	0.3464	0.9432	0.8547
0	1.6	1.9	-1.8	-0.0021	0.3327	0.3418	0.9435	0.0554
0.5	1.6	1.9	-1.8	0.0071	0.3343	0.3373	0.9417	0.3153
1	1.6	1.9	-1.8	-0.0056	0.3350	0.3285	0.9489	0.8477
0	2.3	1.9	-1.8	-0.0056	0.3332	0.3506	0.9313	0.0666
0.5	2.3	1.9	-1.8	0.0015	0.3347	0.3510	0.9399	0.3319
1	2.3	1.9	-1.8	0.0019	0.3332	0.3413	0.9389	0.8462

Table 18: Monte Carlo simulation for heterogeneity of temporal effects by cluster in an expanding cohort SWT design results when analyzing with the the model without random slopes.

				No Random Slope				
Theta	σ_cl	σ_step	σ_cl,step	Bias	Model SE	MC SE	Coverage	Power / Type I error
0	1.6	1.3	-1.8	0.0109	0.8936	3.9347	0.1850	0.8140
0.5	1.6	1.3	-1.8	-0.0186	0.8651	3.7673	0.1810	0.8210
1	1.6	1.3	-1.8	-0.1312	0.8797	3.8073	0.1820	0.7880
0	2.3	1.3	-1.8	0.0705	1.0840	4.9533	0.1590	0.8410
0.5	2.3	1.3	-1.8	-0.0458	1.0582	4.8531	0.1710	0.8240
1	2.3	1.3	-1.8	0.3009	1.0563	4.9033	0.1550	0.8220
0	1.6	1.9	-1.8	-0.1226	0.8701	3.7275	0.1930	0.8070
0.5	1.6	1.9	-1.8	-0.0026	0.8732	3.7456	0.2060	0.7900
1	1.6	1.9	-1.8	0.1876	0.8962	4.0037	0.2130	0.7870
0	2.3	1.9	-1.8	-0.2134	1.0636	4.9060	0.1450	0.8530
0.5	2.3	1.9	-1.8	0.2002	1.1091	5.1874	0.1420	0.8560
1	2.3	1.9	-1.8	-0.4433	1.0808	4.8506	0.1610	0.8080

Table 19: Monte Carlo simulation for heterogeneity of temporal effects by cluster in an expanding cohort SWT design results when analyzing with the the model with neither random slopes nor random intercepts.

				No Random Intercept				
Theta	σ_cl	σ_step	σ_cl,step	Bias	Model SE	MC SE	Coverage	Power / Type I error
0	1.6	1.3	-1.8	0.0160	0.8940	3.5217	0.2220	0.7750
0.5	1.6	1.3	-1.8	-0.0178	0.8653	3.3684	0.2210	0.7840
1	1.6	1.3	-1.8	-0.1205	0.8801	3.3983	0.2230	0.7430
0	2.3	1.3	-1.8	0.0610	1.0853	4.4168	0.1860	0.8140
0.5	2.3	1.3	-1.8	-0.0400	1.0594	4.3349	0.1930	0.7930
1	2.3	1.3	-1.8	0.2627	1.0573	4.3878	0.1880	0.7830
0	1.6	1.9	-1.8	-0.1086	0.8707	3.3358	0.2420	0.7560
0.5	1.6	1.9	-1.8	-0.0035	0.8726	3.3345	0.2400	0.7430
1	1.6	1.9	-1.8	0.1668	0.8971	3.5879	0.2390	0.7550
0	2.3	1.9	-1.8	-0.1797	1.0651	4.3764	0.1770	0.8220
0.5	2.3	1.9	-1.8	0.1924	1.1108	4.6170	0.1730	0.8130
1	2.3	1.9	-1.8	-0.3975	1.0839	4.3335	0.1824	0.7665

$$y_{k(i)j} = (\beta_0 + b_{0i} + b_{0k(i)}) + \beta_1 trt_{ij} + \beta_2 t_j + \beta_3 trt_{ij} t_j + e_{k(i)j} \tag{3.17}$$

for cluster i, time step j, and subject k where

$$b_{0i} \sim N(0, \sigma^2_{cl})$$
$$b_{0k} \sim N(0, \sigma^2_{subj})$$
$$e_{k(ij)} \sim N(0, \sigma^2)$$

To simulate enrollment of subjects after the first step, a multinomial random variable was generated for each subject with probability of enrolling during the first step equaling 0.6 and the probability of enrolling during all later steps having the uniform probability equal to $\frac{0.4}{J-1}$.

Monte Carlo simulations were performed with 1,000 samples using the parameters outlined in Table 20.

Table 20: Simulation parameters for the Monte Carlo study of time-varying treatment effect in an expanding cohort SWT design.

Parameter	Variable	Values
Total number of clusters	I	10
Total number of steps	J	$I+1$
Subjects per cluster per step	K	20
Model intercept	β_0	0.3
Effect of intervention	β_1	$(-1, 0, 0.5, 1)$
Mean time slope	β_2	0.25
Intervention by Time Interaction	β_3	$(-0.10, -0.05, 0, 0.05, 0.10)$
Residual variance	σ^2	1.55^2
Random cluster effect variance	σ^2_{cl}	0.0777^2
Random subject effect variance	σ^2_{subj}	0.1

Each combination of parameters were used to generate $1,000$ Monte Carlo samples. Each simulated data set was then analyzed by two different models: the correct data generating model from Equation (3.17) and a model that omits the intervention by time interaction term as shown in Equation (3.18).

$$y_{k(ij)} = (\beta_0 + b_{0i}) + \beta_1 trt_{ij} + \beta_2 t_j + e_{k(ij)} \tag{3.18}$$

for cluster i, time step j, and subject k where

$$b_{0i} \sim N(0, \sigma^2_{cl})$$
$$e_{k(ij)} \sim N(0, \sigma^2)$$

For each choice of parameters, the mean bias, mean model-based standard error, Monte Carlo standard error, coverage probability, and power/Type I error rate for the effect of intervention will be tested. For the time-varying model, the power/Type I error rate will be reported for an early tstep (i.e., the second step), a middle step ($0.5(J-2)+2$ to the nearest step), and the last step. These results are presented in Tables (21) and (22). For

the time-varying model, the power/Type I error rate will be reported for an early tstep (i.e., the second step), a middle step ($0.5(J-2)+2$ to the nearest step), and the last step. These results are presented in Tables (7) and (8). The power for the contrast tests exhibit a monotone relationship over time when β_1 and $\beta)3$ match in sign. When β_1 and β_3 differ in sign, then . When the signs of β_1 and β_3 differ, the contrast tests show a non-monotonic pattern over time. The time-varying model results show no bias issues. However, the Monte Carlo standard error suggests that the model-based standard error is deflated. The time-fixed treatment effect models do exhibit issues with bias, the sign of which clearly matching the sign of the omitted β_3 interaction term. Monte Carlo standard error suggests that the estimation of the treatment parameter estimate standard error was unbiased. Coverage probability decreases in alignment with the bias issues. Power is decreased especially for the scenarios in which the signs of β_1 and β_3 differ.

3.3.3.3 Cluster-level and Subject-level Temporal Trends To study whether fitting subject-level and cluster-level temporal trends is important for expanding cohort trials, we conducted a Monte Carlo simulation study. Data were generated with 1,000 Monte Carlo samples from expanding cohort trials featuring both cluster-level and subject-level temporal trends. The data were generated according to Equation (3.19), where both of the temporal trends have been fit as parametric linear trends. We chose to focus our study on models with Gaussian-distributed error terms.

$$y_{k(i)j} = \mu + u_i + u_{k(i)} + \theta trt_{ij} + \beta step_j + \tau t_j k + e_{k(i)j} \tag{3.19}$$

$$u_i \sim N(0, \sigma^2_{cl})$$
$$u_{k(i)} \sim N(0, \sigma^2_{subj})$$
$$e_{k(i)j} \sim N(0, \sigma^2)$$

where $t_{jk} = step_j + 1 - t_{0k}$ and t_{0k} is the time at which subject k was recruited into the study.

Time of enrollment into the study was generated from a multinomial distribution with 0.6 probability of enrollment in the first step and probability of $(1.0-0.6)/(J-1)$ for enrollment

Table 21: Monte Carlo simulation results for time-varying treatment effects in an expanding cohort SWT design when analyzing with a time-varying treatment effect model.

					Time-varying			
Beta_1	Beta_3	Bias	Model SE	MC SE	Coverage	Test Early	Test Mid	Test Late
-1	-0.1	-0.0052	0.1137	0.2232	0.9420	1.0000	1.0000	1.0000
-0.5	-0.1	0.0011	0.1138	0.2238	0.9380	0.9930	1.0000	1.0000
0	-0.1	-0.0071	0.1138	0.2196	0.9530	0.2440	1.0000	1.0000
0.5	-0.1	-0.0084	0.1139	0.2238	0.9459	0.4244	0.1832	0.9249
1	-0.1	0.0000	0.1138	0.2317	0.9410	0.9990	0.9690	0.0900
-1	-0.05	0.0128	0.1140	0.2219	0.9480	1.0000	1.0000	1.0000
-0.5	-0.05	-0.0076	0.1139	0.2052	0.9630	0.9620	1.0000	0.9990
0	-0.05	0.0023	0.1139	0.2169	0.9500	0.0920	0.8260	0.8660
0.5	-0.05	0.0031	0.1137	0.2268	0.9440	0.6810	0.4940	0.0580
1	-0.05	0.0075	0.1138	0.2275	0.9389	0.9990	1.0000	0.7184
-1	0	0.0042	0.1139	0.2327	0.9320	1.0000	1.0000	1.0000
-0.5	0	0.0014	0.1136	0.2226	0.9440	0.8460	0.9980	0.7970
0	0	-0.0069	0.1139	0.2203	0.9450	0.0540	0.0510	0.0450
0.5	0	0.0026	0.1137	0.2224	0.9480	0.8490	0.9980	0.7840
1	0	0.0006	0.1139	0.2244	0.9430	1.0000	1.0000	1.0000
-1	0.05	0.0069	0.1138	0.2241	0.9370	1.0000	1.0000	0.7120
-0.5	0.05	-0.0076	0.1138	0.2130	0.9540	0.6790	0.5130	0.0560
0	0.05	0.0003	0.1138	0.2208	0.9490	0.1010	0.8260	0.8790
0.5	0.05	0.0117	0.1137	0.2234	0.9400	0.9410	1.0000	1.0000
1	0.05	0.0012	0.1139	0.2135	0.9570	1.0000	1.0000	1.0000
-1	0.1	0.0074	0.1136	0.2194	0.9480	0.9950	0.9690	0.0810
-0.5	0.1	-0.0103	0.1139	0.2203	0.9530	0.4590	0.1560	0.9140
0	0.1	0.0029	0.1136	0.2158	0.9530	0.2352	1.0000	0.9990
0.5	0.1	-0.0026	0.1139	0.2120	0.9600	0.9910	1.0000	1.0000
1	0.1	0.0066	0.1138	0.2173	0.9460	1.0000	1.0000	1.0000

Table 22: Monte Carlo simulation results for time-varying treatment effects in an expanding cohort SWT design when analyzing with a time-fixed treatment effect model.

			Time-fixed			
Beta_1	**Beta_3**	**Bias**	**Model SE**	**MC SE**	**Coverage**	**Power**
-1	-0.1	-0.6267	0.1039	0.0991	0.0000	1.0000
-0.5	-0.1	-0.6200	0.1037	0.1065	0.0000	1.0000
0	-0.1	-0.6246	0.1038	0.1046	0.0000	1.0000
0.5	-0.1	-0.6279	0.1039	0.1043	0.0000	0.2200
1	-0.1	-0.6202	0.1038	0.1047	0.0010	0.9479
-1	-0.05	-0.3044	0.1035	0.1042	0.1590	1.0000
-0.5	-0.05	-0.3154	0.1033	0.1051	0.1400	1.0000
0	-0.05	-0.3104	0.1033	0.1024	0.1523	0.8427
0.5	-0.05	-0.3116	0.1031	0.1049	0.1470	0.4390
1	-0.05	-0.3098	0.1031	0.1081	0.1520	1.0000
-1	0	-0.0015	0.1032	0.1086	0.9389	1.0000
-0.5	0	0.0007	0.1026	0.1041	0.9380	0.9980
0	0	-0.0055	0.1030	0.1047	0.9469	0.0480
0.5	0	0.0019	0.1029	0.1075	0.9399	0.9980
1	0	-0.0010	0.1030	0.1041	0.9488	1.0000
-1	0.05	0.3083	0.1032	0.1018	0.1411	1.0000
-0.5	0.05	0.3098	0.1032	0.1042	0.1440	0.4560
0	0.05	0.3131	0.1033	0.1033	0.1420	0.8530
0.5	0.05	0.3178	0.1030	0.1063	0.1371	1.0000
1	0.05	0.3077	0.1032	0.1039	0.1471	1.0000
-1	0.1	0.6219	0.1036	0.1056	0.0000	0.9430
-0.5	0.1	0.6158	0.1041	0.1062	0.0000	0.1900
0	0.1	0.6248	0.1038	0.1066	0.0000	1.0000
0.5	0.1	0.6231	0.1039	0.1021	0.0000	1.0000
1	0.1	0.6199	0.1039	0.1052	0.0000	1.0000

in steps $j = 2, 3, \ldots, J$. For this simulation there will be no missing data due to dropout. The total number of clusters I simulated was 10. The number of subjects per cluster K simulated was 5. The linear temporal terms will vary between (-0.01,-0.05,0.05,0.01). The generated data were analyzed with three different models: the true model with both cluster-level and subject-level time trends (see Equation 3.19), a model that only fits cluster-level time trends (see Equation 3.20), and a model that fits only subject-level time trends (see Equation 3.21).

$$y_{k(i)j} = \mu + u_i + u_{k(i)} + \theta trt_{ij} + \beta step_j + e_{k(i)j} \tag{3.20}$$

$$y_{k(i)j} = \mu + u_i + u_{k(i)} + \theta trt_{ij} + \tau t_j k + e_{k(i)j} \tag{3.21}$$

For each Monte Carlo sample, the mean bias, mean model-based standard error, Monte Carlo standard error, coverage probability, and Power/Type I error rate for the effect of intervention is reported in Tables (23), (24) and ,(25).

Table (23) shows the results for estimation and inference of θ (intervention effect) when the generated data where analyzed with the correct model that fits both cluster-level and subject-level trends. As expected this analysis exhibits good performance. The bias is small. The model-based standard errors agree with the Monte Carlo standard errors. The coverage probability remains near 95%. Type I error rate (i.e., power when $\theta = 0$) is well controlled around 5%.

Table (24) shows the results when the data were analyzed with the model as seen in Equation (3.20). Interestingly, the performance of this model is also quite good in terms of bias, standard error estimation, coverage, and control of type I error rate.

Table (25) shows the results when the data were analyzed with the model as seen in Equation (3.21). There is substantial bias in the estimation for θ whenever the underlying cluster trend β is nonzero. As a result, coverage probability decreases well below 95%. Another result is that the Type I error rate is inflated to values between 20% and 30%.

Table 23: Simulation study results when the generated data were analyzed with Equation (3.19).

Both Cluster-level and Subject-level							
θ	τ	β	Bias	Model-based SE	MCMC SE	Coverage	Power
-1	-0.1	-0.1	0.0085	0.241	0.253	0.940	0.978
-0.5	-0.1	-0.1	0.0041	0.241	0.241	0.952	0.532
0	-0.1	-0.1	-0.0089	0.242	0.250	0.940	0.055
-1	0	-0.1	-0.0036	0.241	0.255	0.933	0.978
-0.5	0	-0.1	0.0114	0.241	0.246	0.941	0.518
0	0	-0.1	-0.0043	0.241	0.247	0.941	0.055
-1	0.1	-0.1	0.0082	0.240	0.246	0.944	0.974
-0.5	0.1	-0.1	0.0002	0.240	0.257	0.934	0.534
0	0.1	-0.1	0.0057	0.242	0.258	0.929	0.069
-1	-0.1	0	-0.0091	0.240	0.252	0.938	0.983
-0.5	-0.1	0	0.0063	0.241	0.248	0.945	0.525
0	-0.1	0	0.0041	0.242	0.235	0.951	0.048
-1	0	0	0.0057	0.241	0.247	0.954	0.978
-0.5	0	0	-0.0057	0.240	0.247	0.943	0.538
0	0	0	-0.0032	0.242	0.247	0.931	0.066
-1	0.1	0	-0.0061	0.241	0.245	0.947	0.980
-0.5	0.1	0	-0.0048	0.241	0.248	0.937	0.544
0	0.1	0	0.0058	0.242	0.251	0.942	0.056
-1	-0.1	0.1	0.0003	0.242	0.250	0.940	0.986
-0.5	-0.1	0.1	-0.0173	0.240	0.250	0.941	0.566
0	-0.1	0.1	0.0026	0.241	0.256	0.939	0.058
-1	0	0.1	0.0047	0.240	0.249	0.940	0.981
-0.5	0	0.1	0.0012	0.240	0.264	0.921	0.530
0	0	0.1	0.0032	0.240	0.257	0.933	0.063
-1	0.1	0.1	0.0109	0.241	0.253	0.940	0.977
-0.5	0.1	0.1	0.0110	0.242	0.258	0.926	0.510
0	0.1	0.1	-0.0096	0.242	0.254	0.939	0.058

Table 24: Simulation study results when the generated data were analyzed with Equation (3.20).

Cluster-level Only							
θ	τ	β	Bias	Model SE	MCMC SE	Coverage	Power
-1	-0.1	-0.1	0.0076	0.244	0.254	0.937	0.976
-0.5	-0.1	-0.1	0.0030	0.244	0.245	0.950	0.528
0	-0.1	-0.1	-0.0099	0.245	0.253	0.946	0.052
-1	0	-0.1	-0.0048	0.241	0.255	0.934	0.980
-0.5	0	-0.1	0.0113	0.240	0.246	0.941	0.522
0	0	-0.1	-0.0041	0.241	0.246	0.943	0.057
-1	0.1	-0.1	0.0060	0.243	0.246	0.947	0.974
-0.5	0.1	-0.1	-0.0011	0.243	0.258	0.936	0.524
0	0.1	-0.1	0.0048	0.244	0.257	0.935	0.059
-1	-0.1	0	-0.0082	0.243	0.254	0.938	0.982
-0.5	-0.1	0	0.0060	0.245	0.251	0.940	0.509
0	-0.1	0	0.0050	0.244	0.236	0.951	0.048
-1	0	0	0.0058	0.241	0.246	0.956	0.981
-0.5	0	0	-0.0051	0.240	0.247	0.942	0.541
0	0	0	-0.0027	0.241	0.247	0.934	0.063
-1	0.1	0	-0.0073	0.244	0.246	0.944	0.980
-0.5	0.1	0	-0.0037	0.244	0.250	0.943	0.534
0	0.1	0	0.0033	0.244	0.254	0.937	0.061
-1	-0.1	0.1	0.0021	0.245	0.252	0.936	0.981
-0.5	-0.1	0.1	-0.0178	0.243	0.254	0.941	0.547
0	-0.1	0.1	0.0043	0.244	0.259	0.940	0.059
-1	0	0.1	0.0056	0.240	0.248	0.946	0.982
-0.5	0	0.1	0.0009	0.240	0.263	0.922	0.534
0	0	0.1	0.0021	0.240	0.256	0.933	0.063
-1	0.1	0.1	0.0075	0.244	0.257	0.934	0.974
-0.5	0.1	0.1	0.0090	0.245	0.262	0.933	0.499
0	0.1	0.1	-0.0102	0.244	0.258	0.939	0.057

Table 25: Simulation study results when the generated data were analyzed with Equation (3.21).

Subject-level Only							
θ	τ	β	Bias	Model SE	MCMC SE	Coverage	Power
-1	-0.1	-0.1	-0.2825	0.206	0.220	0.686	1.000
-0.5	-0.1	-0.1	-0.2817	0.207	0.214	0.719	0.965
0	-0.1	-0.1	-0.2910	0.207	0.214	0.703	0.288
-1	0	-0.1	-0.2897	0.206	0.221	0.699	1.000
-0.5	0	-0.1	-0.2806	0.206	0.215	0.711	0.957
0	0	-0.1	-0.2844	0.206	0.213	0.718	0.278
-1	0.1	-0.1	-0.2755	0.206	0.212	0.722	1.000
-0.5	0.1	-0.1	-0.2827	0.206	0.222	0.707	0.955
0	0.1	-0.1	-0.2730	0.207	0.219	0.723	0.272
-1	-0.1	0	-0.0070	0.203	0.211	0.929	0.999
-0.5	-0.1	0	0.0044	0.204	0.208	0.942	0.669
0	-0.1	0	-0.0005	0.204	0.200	0.953	0.042
-1	0	0	0.0102	0.204	0.204	0.953	0.998
-0.5	0	0	-0.0045	0.203	0.205	0.947	0.699
0	0	0	0.0016	0.204	0.210	0.936	0.061
-1	0.1	0	-0.0094	0.203	0.207	0.942	0.997
-0.5	0.1	0	-0.0053	0.204	0.212	0.938	0.678
0	0.1	0	-0.0012	0.204	0.204	0.950	0.048
-1	-0.1	0.1	0.2824	0.207	0.212	0.718	0.923
-0.5	-0.1	0.1	0.2675	0.205	0.218	0.736	0.214
0	-0.1	0.1	0.2847	0.206	0.219	0.700	0.294
-1	0	0.1	0.2904	0.206	0.211	0.717	0.920
-0.5	0	0.1	0.2920	0.206	0.215	0.696	0.169
0	0	0.1	0.2905	0.206	0.221	0.688	0.303
-1	0.1	0.1	0.2895	0.206	0.212	0.700	0.926
-0.5	0.1	0.1	0.2901	0.207	0.216	0.697	0.172
0	0.1	0.1	0.2745	0.207	0.213	0.731	0.266

3.4 DISCUSSION

SWT designs can vary greatly in terms of the design and may be subject to various assumptions about the underlying data. We conducted Monte Carlo simulations to study three assumptions (i.e., heterogeneity of temporal trends, time-varying treatment effect, and cluster-level versus subject-level temporal trends). Out simulation studies chose to focus on model with Gaussian-distributed error terms. However, these assumptions could apply just as well to a generalized linear mixed model for binary or count outcomes. Heterogeneity of temporal trends and time-varying treatment effects were studied in all three design types (i.e., cross-sectional, fixed cohort, and expanding cohort). Cluster-level and subject-level temporal trends were studied only for expanding cohort designs since they were the only design for which step indicator variables might not properly control for both trends.

Our simulation study on with heterogeneous temporal effects by cluster shows that failure to account for that heterogeneity in the analysis model can lead to biased estimation of the standard error for the intervention effect parameter. This issue leads to poor coverage probability and inflated type I error rate for the intervention effect. We chose to study the temporal trends as a parametric linear form. Future work could examine different forms of temporal trends. An interesting topic for future research would be to simulate data in which multiple clusters are randomized to a crossover time together. Such randomization might help reduce the conflation between cluster-specific temporal effects and the intervention effect.

When the effect of intervention is time-varying, failure to incorporate that variation in the model will result in poor estimation of the intervention effect. For our study, we chose a parametric linear form for time. A nonparametric temporal trend can be fit by using indicator variables for the steps. To be more general, one could chose to fit interaction terms between the effect of intervention and step indicator terms. In many settings, the investigators are interested in testing for a main effect of intervention. There are a couple published options for accurately finding some main effect of intervention in the presence of time-varying effect. Hussey and Hughes[10] recommended allowing the values of the design matrix for treatment

effect (i.e., trt_{ij}) to be between 0 and 1 to represent incomplete attainment of the full intervention effect.

In later publication, Hughes et al.[9] suggested a data-driven two-step procedure for estimating a long-term intervention effect when the effect is time-varying. First, a model is fit including a term $L_{ijl}\theta_l$, where L_{ijl} is an indicator for whether cluster i was on treatment for l steps by step j and θ_l are the parameters for the effect of being on intervention for l steps. Then, the $\hat{\theta}_l$ estimates from model are substituted into an explicit model relating θ_l to a long-term effect of intervention θ_0 using nonlinear weighted least squares with weights provided from the covariance matrix for $\hat{\theta}_l$. Standard errors for the parameters can be obtained via bootstrap.

For expanding cohort designs, our simulation study suggests that not fitting subject-level temporal trends has virtually no consequences for the estimation and inference for the effect of intervention. Thus, expanding cohort studies should yield reasonable estimation and inference for intervention effect using the same analysis model as for a fixed cohort study. The main difference being that the expanding cohort design has unbalanced measurements due to subjects not having values prior to recruitment. For this reason, procedures that rely on balanced and complete data such as GEE should be avoided. For our study, the temporal trends were both assumed to be linear. Future work could examine whether our results hold for other forms of temporal trends.

4.0 ANALYTICAL ASSUMPTIONS FOR STEPPED WEDGE CLUSTER RANDOMIZED TRIALS

ABSTRACT

Stepped wedge cluster randomized trials are a type of cluster randomized trial involving repeated measures over time with unidirectional crossover from control to intervention. In this paper, we focus on two aspects to the analysis of SWT designs: choice of correlation structure and fitting of the step effect. We conducted Monte Carlo simulation studies to determine how these analytical assumptions affect the estimation and inference for the effect of intervention. Both analytical issues were studied in each of the three types of SWT designs: cross-sectional, fixed cohort, and expanding cohort. For choice of correlation structure, we found that choosing an exchangeable correlation structure when the true data were generated as first order autoregressive resulted in deflated estimates of the standard error for intervention effect. These deflated estimates can result in inflated Type I error rate. For the fixed and expanding cohort designs, choosing the first order autoregressive stucture when the data were generated as exchangeable resulted in inflated estimates of the standard error for the intervention effect. These inflated estimates can result in an overly conservative Type I error rate. Analysis using generalized estimating equations (GEE) resulted in deflated standard error estimates when using the empirical variance estimator. This deflation was not affected by misspecification of the correlation structure for cross-sectional sWT designs.

For fixed cohort designs, misspecification of the correlation structure with GEE had similar results except that Type I error what better controlled when the underlying data were exchangeable and the working correlation structure was chosen as exchangeable. Our simulation studies on choice of fitting the step effect as fixed versus random suggest that this choice does not influence the estimation or inference of the intervention effect substantially.

Stepped wedge trial (SWT) designs are commonly used in the setting of public health and policy. Ensuring that SWTs are adequately powered and correctly analyzed will have a beneficial impact on research for public health.

KEY WORDS : Cluster randomized trial,stepped wedge trial,Monte Carlo simulation.

4.1 INTRODUCTION

Stepped wedge trial (SWT) designs are a type of cluster randomized trial in which clusters are randomized to a pre-determined time point (i.e. step) during which they switch from control to intervention. These trials are commonly employed for community-based health research for public health and public policy. In a previous publication[cite first paper], we proposed three terms to classify subtypes of SWT designs. These types are cross-sectional, fixed cohort, and expanding cohort. Cross-sectional SWT designs are defined by subjects being included for only one step of exposure and being recruited throughout the study. Fixed cohort SWT designs are defined by subjects being recruited completely at the beginning of the study and then followed until the end of the study or until dropout. Expanding cohort SWT designs are defined by subjects being recruited throughout the study (though usually many are recruited at the start) and followed until either the end of the study or dropout. In this paper, we conducted Monte Carlo simulation studies on the two analytical assumptions that we discussed in [first paper cite]. For each of the three different types of SWT designs, we conduct a separate Monte Carlo simulation study on the choice of correlation structure and the fitting of the step effect on the estimation and inference for the effect of intervention. This paper is organized as follows. In section 4.2, we discuss the notations and models to provide background on choice of correlation structure and fitting of the step effect. In Section 4.3, we present the Monte Carlo simulation study setup and results. In section 4.4, we discuss the results of our simulation studies and what they entail for the analysis of SWT designs.

4.2 NOTATION AND MODELS

4.2.1 Choice of correlation structure

Hussey and Hughes [10] suggested a statistical model for analyzing cross-sectional SWTs using a linear mixed model (LMM) with random effects for clusters. Another approach that has been used for analysis of SWTs is generalized estimating equations (GEE). GEE has the advantage that it provides results that are robust to misspecification of the working correlation structure. [13] However, one drawback to GEE is that it only uses complete cases. For that reason GEE might provide biased estimates in the presence of missing data that are not missing completely at random (MCAR). Conversely, LMMs yield results that are sensitive to the choice of correlation structure. Lastly, the empirical variance estimator commonly used for GEE relies on asymptotic properties; thus, it might not provide accurate inference when there are few clusters in the study. Alternatives to the empirical variance estimator for GEE are available, though. Scott et al. [18] studied the use of various finite-sample corrected variance estimates for GEE in SWT designs. Despite these facts, LMM remains the most popular methods for analyzing SWTs.

Since LMM and GLMM remain one of the more popular modeling strategies for SWT designs, we conducted Monte Carlo simulation studies to study how choice of correlation structure affects the estimation and inference of intervention effects. This study focuses on exchangeable and first-order autoregressive correlation structures. For SWT designs without repeated measures on subjects, only within-cluster correlation needs to be considered. Other designs need to consider both the within-cluster correlation and the within-subject correlation. Each simulated set will be analyzed both the correct and incorrect correlation structures.

4.2.2 Fixed versus random effects for step

In Hussey and Hughes [10], they recommended fitting indicator variables for the step in order to account for secular trends. The authors made a brief comment about a suggestion from reviewers for whether the step effect could be fit as random effects instead. They replied that

"We felt that this approach did not reflect our interest in controlling for temporal trends and fluctuations in disease prevalence over the course of a particular trial" [10]. Nonetheless, they did suggest that random effects for step might be appropriate in other settings and was worth further investigation. In this study, we conducted Monte Carlo simulation studies to examine how including a random effect term for the step influences estimation and inference for the effect of intervention. We consider three scenarios for data generation: step effects that are random with a nonzero mean, the steps are fixed effects, and an interaction term between step and intervention.

4.3 SIMULATION STUDIES

This section will focus on analytical issues that arise for SWT designs. This section is divided first by type of SWT design (open cross-sectional, fixed cohort, and expanding cohort) then by analytical issues for each SWT design type. Specifically, the analytical issues presented here are the choice of correlation structure and the choice of fitting the step effects as fixed or random effects.

4.3.1 Cross-sectional SWT Designs

4.3.1.1 Choice of Correlation Structure Monte Carlo simulation studies were conducted to study the effect that choice of correlation structure (and misspecification) had on the estimation and inference for the effect of intervention. Since in the case of cross-sectional SWT designs exchangeable correlation structure is the only sensible choice for subject-level models, this simulation study was conducted with a cluster-level model.

$$y_{ij} = \beta_0 + \beta_1 trt_{ij} + \beta_2 t_j + e_{ij} \tag{4.1}$$

for cluster i, time step j

Table 26: Simulation parameters for the Monte Carlo study on choice of correlation structure for cross-sectional SWT designs.

Parameter	Variable	Values
Total number of clusters	I	20
Total number of steps	J	$I+1$
Model intercept	β_0	0.3
Effect of intervention	β_1	$(0, 0.5, 1)$
Step Effect	β_2	0.25
Residual variance	σ^2	1.55^2
Correlation parameter	ρ	0.6

Correlation structures to account for within-cluster correlation will be either exchangeable or autoregressive first order. The parameters for the simulation were chosen as shown in Table (26).

Each simulated data set will be analyzed using linear mixed models with an exchangeable correlation structure and autoregressive first order correlation structure. Additionally, each Monte Carlo sample will be analyzed using generalized estimating equation with an exchangeable correlation structure. For each analysis, the bias, mean model-based (or empirical or GEE) standard error, Monte Carlo standard error, coverage probability, and power/Type I error rate for the effect of intervention are reported in Table (26).

Table (27) shows that there is no issue with bias in the intervention effect estimation. However, the estimation of the standard error for the intervention effect is incorrectly estimated when compared to the Monte Carlo standard error is some of the following situations. When the underlying correlation structure is autoregressive, but the data are analyzed as exchangeable, the estimated standard error is deflated resulting in loss of coverage probability and inflation of Type I error rate. Conversely, when the underlying correlation pattern is exchangeable, but the data are analyzed as autoregressive, the standard error estimate corresponds well with the Monte Carlo standard error. When the generated data were analyzed

Table 27: Results from the Monte Carlo simulation study for a cluster-level model for a cross-sectional SWT design.

θ	Correlation Generated	Analysis Correlation	Bias	MC Standard Error	Estimated Standard Error	Coverage	Power
0	EXCH	EXCH	-0.0024	0.2601	0.2584	0.9510	0.0490
0.5	EXCH	EXCH	-0.0057	0.2526	0.2584	0.9520	0.4790
0	AR	EXCH	0.0026	0.4262	0.2739	0.7950	0.2010
0.5	AR	EXCH	0.0094	0.4134	0.2736	0.7940	0.4710
0	EXCH	AR	-0.0020	0.2603	0.2581	0.9500	0.0500
0.5	EXCH	AR	-0.0053	0.2535	0.2580	0.9450	0.4820
0	AR	AR	0.0051	0.3261	0.3295	0.9500	0.0500
0.5	AR	AR	0.0068	0.3223	0.3292	0.9550	0.3450
0	EXCH	GEE (EXCH)	-0.0024	0.2602	0.2482	0.9280	0.0720
0.5	EXCH	GEE (EXCH)	-0.0058	0.2527	0.2476	0.9360	0.5100
0	AR	GEE (EXCH)	0.0029	0.4242	0.3968	0.9260	0.0740
0.5	AR	GEE (EXCH)	0.0089	0.4117	0.3979	0.9340	0.2730
0	EXCH	GEE (AR)	-0.0063	0.4859	0.4916	0.9460	0.0540
0.5	EXCH	GEE (AR)	-0.0246	0.4832	0.4847	0.9250	0.1810
0	AR	GEE (AR)	0.0056	0.3277	0.3136	0.9250	0.0750
0.5	AR	GEE (AR)	0.0070	0.3230	0.3145	0.9430	0.3650

Theta is the intervention effect parameter. The abbreviations 'EXCH' and 'AR' stand for exchangeable and autoregressive respectively. For GEE analyses, the estimated standard error is based on the empirical variance estimator. The MC error represents the Monte Carlo standard error. Note that when theta is zero, 'Power' refers to Type I error rate.

using GEE, the empirical estimate for the standard error is is only very slightly deflated compared to the Monte Carlo standard error. This slight deflation of the standard error is not further affected by misspecification of the correlation structure. This observation suggests that the slight deflation is due to the relatively small sample size of 20 clusters since the empirical variance estimator relies on asymptotic properties.

4.3.1.2 Fixed versus Random Effects for Steps We conducted Monte Carlo simulations to study how fitting random effects for steps influences the estimation and inference for the intervention effect under three different data generating models. First, we examine the case when data were generated with both random and fixed effects for the step effect. These data were analyzed with both the true model as well as a model that only fits the fixed effect. Secondly, we consider the case when the data were generated with fixed effects for the step effect. These data were then analyzed both by the true model as well as a model that fits step as mean zero random effects rather fixed effects. Lastly, we conducted a simulation to study in which the data were generated with an interaction term between time and intervention. These data were analyzed both with the true model as well as a model that fits steps as random effects and omits the interaction term.

For the first simulation studying the use of random effects for step, data was generated according to the model outlined in the equation below.

$$y_{k(ij)} = (\mu + u_i + u_j) + \theta trt_{ij} + \beta_j + e_{k(ij)} \tag{4.2}$$

for cluster i, time step j, and subject k where

$$u_i \sim N(0, \sigma^2_{cl})$$
$$u_j \sim N(0, \sigma^2_{step})$$
$$e_{k(ij)} \sim N(0, \sigma^2)$$

Table 28: Simulation parameters for the Monte Carlo study on fitting the step effect as fixed or random in a cross-sectional SWT design.

Parameter	Variable	Values
Total number of clusters	I	10
Total number of steps	J	$I+1$
Subjects per cluster per step	K	20
Model intercept	μ	0.3
Effect of intervention	β_1	$(0, 0.5, 1)$
Mean step effects	β_j	$\beta_j = (0.25(j-1))$ for $j = 2, \ldots, J$
Residual variance	σ^2	1.55^2
Random cluster effect variance	σ^2_{cl}	0.0777^2
Random step effect variance	σ^2_{step}	$(0, 0.0777^2, 0.157^2)$

All simulations shared the parameters outlined in Table 28. Note that the mean step effects were chosen to fit a linear trend. One thousand Monte Carlo samples were generated for each choice of parameters. For each sample, the data was analyzed both by fitting the step effect and random and by only fitting the step effect as fixed. For each analysis, the bias, mean model standard error, Monte Carlo standard error, coverage probability, and power/Type I error rate for the effect of intervention are reported in Tables (29) and (30). Simulation results do not suggest any differences in estimation or inference for the treatment effect based on fitting the step effect as random or fixed.

For the second random step simulation study, we generated data according to equation (4.3) below.

$$y_{k(ij)} = \mu + u_i + \theta trt_{ij} + \beta t_j + e_{k(ij)} \tag{4.3}$$

for cluster i, time step j, and subject k

$$u_i \sim N(0, \sigma^2_{cl})$$
$$e_{k(ij)} \sim N(0, \sigma^2)$$

Table 29: Monte Carlo simulation study results for fitting the step effect as fixed or random in an open cross-sectional SWT design.

		Random Step Model				
Theta	sigma_step	Bias	Model SE	MC SE	Coverage	Power/Type I Error Rate
0	0	-0.0045	0.0918	0.0986	0.9340	0.0630
0.5	0	-0.0071	0.0921	0.0971	0.9370	0.9990
1	0	0.0054	0.0922	0.0953	0.9370	1.0000
0	0.0777	-0.0048	0.0921	0.0958	0.9420	0.0560
0.5	0.0777	-0.0051	0.0922	0.0955	0.9510	1.0000
1	0.0777	-0.0014	0.0925	0.0965	0.9360	1.0000
0	0.157	-0.0062	0.0919	0.0979	0.9240	0.0730
0.5	0.157	0.0022	0.0921	0.0943	0.9440	1.0000
1	0.157	-0.0010	0.0921	0.0939	0.9440	1.0000

These results are when the data were analyzed fitting the step effect as random.

Table 30: Monte Carlo simulation study results for fitting the step effect as fixed or random in an open cross-sectional SWT design.

		Fixed Step Model				
Theta	sigma_step	Bias	Model SE	MC SE	Coverage	Power/Type I Error Rate
0	0	-0.0047	0.0918	0.0986	0.9339	0.0631
0.5	0	-0.0071	0.0921	0.0971	0.9370	0.9990
1	0	0.0053	0.0922	0.0954	0.9369	1.0000
0	0.0777	-0.0048	0.0922	0.0958	0.9420	0.0560
0.5	0.0777	-0.0051	0.0923	0.0955	0.9510	1.0000
1	0.0777	-0.0014	0.0926	0.0965	0.9370	1.0000
0	0.157	-0.0062	0.0921	0.0978	0.9250	0.0720
0.5	0.157	0.0024	0.0924	0.0943	0.9449	1.0000
1	0.157	-0.0010	0.0924	0.0939	0.9450	1.0000

These results are when the data were analyzed fitting the step effect as fixed.

Table 31: Simulation parameters for the Monte Carlo study on fitting the step effect as fixed or mean zero random effect in a cross-sectional SWT design.

Parameter	Variable	Values
Total number of clusters	I	10
Total number of steps	J	$I+1$
Subjects per cluster per step	K	10
Model intercept	μ	0.3
Effect of intervention	θ	$(0, 0.5, 1)$
Mean step effects	β	$\beta = (-0.2, -0.1, 0, 0.1, 0.2)$
Residual variance	σ^2	1.55^2
Random cluster effect variance	σ^2_{cl}	0.0777^2

In addition to analyzing these data with the correct model, we also analyzed them using a model that fits the step effects as mean zero random effects (See Equation 4.4).

$$y_{k(ij)} = (\mu + u_i + u_j) + \theta trt_{ij} + e_{k(ij)} \tag{4.4}$$

for cluster i, time step j, and subject k where

$$u_i \sim N(0, \sigma^2_{cl})$$
$$u_j \sim N(0, \sigma^2_{step})$$
$$e_{k(ij)} \sim N(0, \sigma^2)$$

The simulations were run with 1000 Monte Carlo samples using the parameters outlined in Table (31).

For each analysis, the bias, mean model standard error, Monte Carlo standard error, coverage probability, and power/Type I error rate for the effect of intervention are reported in Table (32). As expected, when analyzing the data with the correct model using a fixed effect for step, the estimation and inference for the intervention effect are both appropriate. However, when the data were analyzed using a mean zero random effect for step, there is bias in

the estimation and the model-based standard error estimates are deflated compared to the Monte Carlo standard error. It appears that both the estimation and inference issues are actually worse for the smaller magnitude of the step slope (i.e., $|\beta| = 0.1$). When there was no true intervention effect (i.e., $\theta = 0$), these issues result in inflated Type I error rate when $\beta \neq 0$.

Lastly, we conducted a simulation study in which data were generated with an interaction term between intervention effect and step. Data were generated according to the Equation (4.5).

$$y_{k(ij)} = \mu + u_i + \theta trt_{ij} + \beta t_j + \gamma t_j trt_{ij} + e_{k(ij)} \tag{4.5}$$

for cluster i, time step j, and subject k where

$$u_i \sim N(0, \sigma^2_{cl})$$
$$e_{k(ij)} \sim N(0, \sigma^2)$$

In addition to analyzing these data with the correct model, we also analyzed them using a model that fits mean zero random effects for steps without including the interaction term between step and intervention (See Equation 4.6).

$$y_{k(ij)} = (\mu + u_i + u_j) + \theta trt_{ij} + \beta t_j + e_{k(ij)} \tag{4.6}$$

for cluster i, time step j, and subject k where

$$u_i \sim N(0, \sigma^2_{cl})$$
$$u_j \sim N(0, \sigma^2_{step})$$
$$e_{k(ij)} \sim N(0, \sigma^2)$$

The simulations were run with 1000 Monte Carlo samples using the parameters outlined in Table (33).

For each analysis, the bias, mean model standard error, Monte Carlo standard error, coverage probability, and power/Type I error rate for the effect of intervention are reported in Table (33). As expected, when analyzing the data with the correct model using a fixed effect for

Table 32: Monte Carlo simulation study results for fitting the step effect as fixed or random in a cross-sectional SWT design.

Cross-sectional SWT Design							
θ	β1	Step Effect	Bias	Model SE	MC SE	Coverage	Power
0	0	Random	0.0025	0.0988	0.1017	0.9400	0.0550
0.5	0	Random	0.0053	0.0986	0.0961	0.9560	0.9990
1	0	Random	0.0041	0.0988	0.0974	0.9540	1.0000
0	0.1	Random	0.1722	0.1163	0.1478	0.6410	0.3470
0.5	0.1	Random	0.1732	0.1167	0.1528	0.6350	1.0000
1	0.1	Random	0.1534	0.1168	0.1484	0.6907	1.0000
0	0.2	Random	0.0808	0.1252	0.1332	0.8740	0.1220
0.5	0.2	Random	0.0809	0.1246	0.1389	0.8710	0.9950
1	0.2	Random	0.0854	0.1248	0.1394	0.8619	1.0000
0	-0.1	Random	-0.1610	0.1170	0.1370	0.7010	0.2980
0.5	-0.1	Random	-0.1663	0.1169	0.1451	0.6830	0.7680
1	-0.1	Random	-0.1583	0.1169	0.1505	0.6770	1.0000
0	-0.2	Random	-0.0874	0.1248	0.1356	0.8770	0.1190
0.5	-0.2	Random	-0.0837	0.1249	0.1355	0.8700	0.8820
1	-0.2	Random	-0.0797	0.1249	0.1379	0.8719	1.0000
0	0	Fixed	0.0062	0.1272	0.1279	0.9479	0.0501
0.5	0	Fixed	0.0091	0.1271	0.1262	0.9489	0.9820
1	0	Fixed	0.0097	0.1269	0.1270	0.9440	1.0000
0	0.1	Fixed	0.0086	0.1272	0.1257	0.9470	0.0500
0.5	0.1	Fixed	0.0093	0.1275	0.1314	0.9330	0.9690
1	0.1	Fixed	-0.0042	0.1269	0.1301	0.9419	1.0000
0	0.2	Fixed	-0.0049	0.1275	0.1279	0.9418	0.0542
0.5	0.2	Fixed	-0.0031	0.1270	0.1328	0.9499	0.9679
1	0.2	Fixed	0.0008	0.1272	0.1329	0.9340	1.0000
0	-0.1	Fixed	-0.0015	0.1273	0.1193	0.9659	0.0311
0.5	-0.1	Fixed	-0.0035	0.1274	0.1230	0.9530	0.9720
1	-0.1	Fixed	0.0025	0.1274	0.1291	0.9420	1.0000
0	-0.2	Fixed	-0.0026	0.1272	0.1286	0.9480	0.0510
0.5	-0.2	Fixed	0.0014	0.1272	0.1292	0.9420	0.9690
1	-0.2	Fixed	0.0045	0.1272	0.1314	0.9349	1.0000

The data were generated with a linear parametric term for step. The parameters θ and β represent the intervention effect and the linear step slope respectively. The 'Step Effect' column indicates whether the data were analyzed using a fixed effect (See Equation 4.3) or a mean zero random effect (See Equation 4.4). The bias, average model standard error, Monte Carlo standard error, coverage probability, and power are presented with respect to the effect of intervention parameter θ. Note that when $\theta = 0$, power actually refers to Type I error rate.

Table 33: Simulation parameters for the Monte Carlo study on fitting a random step effect in place of an interaction term between step and intervention effect in a cross-sectional SWT design.

Parameter	Variable	Values
Total number of clusters	I	10
Total number of steps	J	$I+1$
Subjects per cluster per step	K	10
Model intercept	μ	0.3
Effect of intervention	θ	$(0, 0.5, 1)$
Intervention and step interaction	γ	$(-0.05, -0.02, 0, 0.02, 0.05)$
Step slope parameter	β	$\beta = 0.25$
Residual variance	σ^2	1.55^2
Random cluster effect variance	σ^2_{cl}	0.0777^2

step, the estimation and inference for the intervention effect are both appropriate. When analyzing the data using the model with a random step effect, there is a small deflation of the standard error estimates and bias in the estimate for the main intervention effect. These issues result in inflation of Type I error rate even when the interaction term was zero. Also, for this analysis model, the coverage probability decreases for larger absolute values of the interaction term.

4.3.2 Fixed cohort SWT Designs

4.3.2.1 Choice of Correlation Structure Monte Carlo simulation studies were conducted to study the effect that choice of correlation structure (and misspecification) had on the estimation and inference for the effect of intervention. Data were generated under the mean model shown in equation (4.7). Within-cluster correlation was modeled using random effects. Within-subject correlation was modeled through a covariance pattern with either an exchangeable or autoregressive structure.

Table 34: This table summarizes the Monte Carlo simulation study results for fitting the step as a random effect in a cross-sectional SWT design with an underlying interaction term between step and intervention.

Cross-sectional SWT							
θ	γ	Model	Bias	Model SE	MC SE	Coverage	Power
0	0	Random effect	0.0720	0.1258	0.1319	0.9080	0.0900
0.5	0	Random effect	0.0729	0.1256	0.1345	0.8980	0.9920
1	0	Random effect	0.0608	0.1256	0.1331	0.9090	1.0000
0	0.02	Random effect	0.1807	0.1260	0.1317	0.6810	0.3100
0.5	0.02	Random effect	0.1840	0.1255	0.1328	0.6800	0.9990
1	0.02	Random effect	0.1855	0.1260	0.1318	0.6750	1.0000
0	0.05	Random effect	0.3724	0.1261	0.1419	0.1900	0.8040
0.5	0.05	Random effect	0.3674	0.1263	0.1341	0.1710	1.0000
1	0.05	Random effect	0.3588	0.1264	0.1367	0.2160	1.0000
0	-0.02	Random effect	-0.0503	0.1253	0.1364	0.9020	0.0970
0.5	-0.02	Random effect	-0.0456	0.1253	0.1327	0.9190	0.9300
1	-0.02	Random effect	-0.0484	0.1257	0.1353	0.9090	1.0000
0	-0.05	Random effect	-0.2250	0.1257	0.1346	0.5460	0.4490
0.5	-0.05	Random effect	-0.2320	0.1256	0.1318	0.5290	0.5380
1	-0.05	Random effect	-0.2275	0.1257	0.1331	0.5530	1.0000
0	0	Interaction term	0.0087	0.2634	0.2735	0.9330	0.0670
0.5	0	Interaction term	-0.0028	0.2632	0.2607	0.9570	0.4640
1	0	Interaction term	-0.0097	0.2632	0.2602	0.9450	0.9600
0	0.02	Interaction term	0.0017	0.2633	0.2665	0.9520	0.0480
0.5	0.02	Interaction term	-0.0090	0.2631	0.2655	0.9449	0.4635
1	0.02	Interaction term	-0.0060	0.2632	0.2622	0.9540	0.9690
0	0.05	Interaction term	0.0130	0.2632	0.2676	0.9419	0.0561
0.5	0.05	Interaction term	0.0035	0.2635	0.2690	0.9370	0.4800
1	0.05	Interaction term	0.0137	0.2637	0.2635	0.9510	0.9720
0	-0.02	Interaction term	0.0001	0.2629	0.2742	0.9440	0.0560
0.5	-0.02	Interaction term	0.0194	0.2632	0.2749	0.9350	0.5150
1	-0.02	Interaction term	-0.0001	0.2634	0.2558	0.9540	0.9680
0	-0.05	Interaction term	0.0045	0.2634	0.2700	0.9510	0.0490
0.5	-0.05	Interaction term	-0.0016	0.2632	0.2694	0.9420	0.4750
1	-0.05	Interaction term	-0.0061	0.2635	0.2587	0.9560	0.9670

The data were generated with a linear parametric term for step. The parameters θ and γ represent the main intervention effect and the step by intervention interaction term respectively. The 'Model' column indicates whether the data were analyzed using the data generating model with an the interaction between step and intervention (See Equation 4.5) or a mean zero random effect (See Equation 4.6). The bias, average model standard error, Monte Carlo standard error, coverage probability, and power are presented with respect to the main effect of intervention parameter θ. Note that when $\theta = 0$, power actually refers to Type I error rate.

Table 35: Simulation parameters for the Monte Carlo study on choice of correlation structure for fixed cohort SWT designs.

Parameter	Variable	Values
Total number of clusters	I	10
Total number of steps	J	$I+1$
Subjects per cluster per step	K	20
Model intercept	β_0	0.3
Effect of intervention	β_1	$(0, 0.5)$
Step Effect	β_2	0.25
Residual variance	σ^2	1.55^2
Random cluster effect variance	σ^2_{cl}	1.20
Correlation parameter	ρ	0.6

The parameters for the simulation were chosen as shown in Table (35).

$$y_{k(i)j} = \beta_0 + u_i + u_{k(i)} + \beta_1 trt_{ij} + \beta_2 t_j + e_{k(i)j} \tag{4.7}$$

for cluster i, time step j, and subject k

where $u_i \sim N(0, \sigma^2_{cl})$ are the random effects for clusters and $u_{k(i)} \sim N(0, \sigma^2_{subj})$.

Each simulated data set will be analyzed using linear mixed models with that fit random effects for the clusters and use a covariance pattern to handle within-subject correlation. Additionally, each Monte Carlo sample will be analyzed using generalized estimating equation with an exchangeable correlation structure both at the cluster level and at the subject level. For each analysis, the bias, mean model-based (or empirical or GEE) standard error, Monte Carlo standard error, coverage probability, and power/Type I error rate for the effect of intervention are reported in Table (36).

Table 36: Results from the Monte Carlo simulation study for a fixed cohort SWT design. Theta is the intervention effect parameter.

θ	Correlation Generated	Analysis Correlation	Bias	MC Standard Error	Estimated Standard Error	Coverage	Power
0.5	AR	AR	0.0013	0.1301	0.1300	0.9490	0.9710
0.5	AR	EXCH	-0.0019	0.1668	0.1261	0.8610	0.9390
0.5	AR	GEE (AR)	0.0003	0.1855	0.1261	0.8010	0.9110
0.5	AR	GEE (Cluster-level)	-0.0020	0.1670	0.1491	0.8960	0.8690
0.5	AR	GEE (EXCH)	-0.0032	0.2035	0.1569	0.8630	0.8210
0	AR	AR	-0.0011	0.1249	0.1299	0.9630	0.0350
0	AR	EXCH	-0.0036	0.1609	0.1259	0.8730	0.1270
0	AR	GEE (AR)	-0.0036	0.1759	0.1260	0.8420	0.1580
0	AR	GEE (Cluster-level)	-0.0037	0.1611	0.1474	0.8980	0.1020
0	AR	GEE (EXCH)	-0.0068	0.1981	0.1565	0.8790	0.1210
0.5	EXCH	AR	-0.0003	0.1470	0.1640	0.9700	0.8860
0.5	EXCH	EXCH	-0.0005	0.1137	0.1153	0.9490	0.9920
0.5	EXCH	GEE (AR)	-0.0037	0.2099	0.1547	0.8460	0.8180
0.5	EXCH	GEE (Cluster-level)	-0.0009	0.1158	0.1064	0.8970	0.9860
0.5	EXCH	GEE (EXCH)	-0.0014	0.1224	0.1130	0.9130	0.9860
0	EXCH	AR	-0.0035	0.1452	0.1640	0.9760	0.0240
0	EXCH	EXCH	-0.0011	0.1142	0.1154	0.9470	0.0520
0	EXCH	GEE (AR)	0.0007	0.2057	0.1547	0.8520	0.1480
0	EXCH	GEE (Cluster-level)	-0.0008	0.1152	0.1059	0.9060	0.0940
0	EXCH	GEE (EXCH)	0.0001	0.1200	0.1132	0.9350	0.0650

The abbreviations 'EXCH' and 'AR' stand for exchangeable and autoregressive respectively. The MC error represents the Monte Carlo standard error. Note that when theta is zero, 'Power' refers to Type I error rate.

In Table (36), one can see that when the correct correlation structure is chosen for a LMM, the bias, standard error estimation, coverage, and power exhibit good properties. When the data are generated as autoregressive, but analyzed as exchangeable, the standard error estimate is deflated compared to the Monte Carlo standard error. Converesly, when the data are generated as exchangeable, but are analyzed as autoregressive, the standard error estimate is inflated compared to the Monte Carlo standard error. In terms of Type I error rate, this results in inflated Type I error rate when the true underlying correlation structure is autoregressive, but fit as exchangeable and deflated (conservative) Type I error rate when the true underlying correlation structure is exchangeable, but fit as autoregressive. When using GEE to analyze, the empirical estimator for the standard error is deflated compared to the Monte Carlo standard error. This observation is expected since it relies on asymptotic theory. In general, GEE at the cluster level using exchangeable correlation yields better inference than either GEE-based subject-level analysis.

4.3.2.2 Fixed versus Random Effects for Steps For the first simulation studying the use of random effects for step, data was generated according to the model outlined in Equation (4.8) below.

$$y_{k(ij)} = (\mu + u_i + u_j + u_{k(i)}) + \beta_1 trt_{ij} + \beta_j + e_{k(ij)} \tag{4.8}$$

for cluster i, time step j, and subject k where

$$u_i \sim N(0, \sigma^2_{cl})$$
$$u_j \sim N(0, \sigma^2_{step})$$
$$u_{k(i)} \sim N(0, \sigma^2_{subj}) \; e_{k(ij)} \sim N(0, \sigma^2)$$

All simulations shared the parameters outlined in Table 37. Note that the mean step effects were chosen to fit a linear trend.

One thousand Monte Carlo samples were generated for each choice of parameters. For each sample, the data was analyzed both by fitting the step effect and random and by only fitting the step effect as fixed. For each analysis, the bias, mean model standard error, Monte

Table 37: Simulation parameters for the Monte Carlo study on fitting the step effect as fixed or random in a fixed cohort SWT design.

Parameter	Variable	Values
Total number of clusters	I	10
Total number of steps	J	$I+1$
Subjects per cluster	K	20
Model intercept	μ	0.3
Effect of intervention	β_1	$(0, 0.5, 1)$
Mean step effects	β_j	$\beta_j = (0.25(j-1))$ for $j = 2, \ldots, J$
Residual variance	σ^2	1.55^2
Random cluster effect variance	σ^2_{cl}	0.0777^2
Random step effect variance	σ^2_{step}	$(0, 0.0777^2, 0.157^2)$
Random subject effect variance	σ^2_{subj}	0.1

Carlo standard error, coverage probability, and power/Type I error rate for the effect of intervention are reported in Tables (38) and (39). Simulation results do not suggest any differences in estimation or inference for the treatment effect based on fitting the step effect as random or fixed.

For the second random step simulation study, we generated data according to equation (4.9) below.

$$y_{k(i)j} = (\mu + u_i + u_k(i)) + \theta trt_{ij} + \beta t_j + e_{k(i)j} \tag{4.9}$$

for cluster i, time step j, and subject k

$$u_i \sim N(0, \sigma^2_{cl})$$
$$u_k(i) \sim N(0, \sigma^2_{subj})$$
$$e_{k(i)j} \sim N(0, \sigma^2)$$

Table 38: Monte Carlo simulation study results for fitting the step effect as fixed or random in a fixed cohort SWT design. These results are when the data were analyzed fitting the step effect as random.

			Random Step Model				
Theta	Sigma_subj	sigma_step	Bias	Model SE	MC SE	Coverage	Power/Type I Error Rate
0	0	0	-0.0002	0.0918	0.0957	0.9419	0.0561
0.5	0	0	0.0025	0.0922	0.0959	0.9389	1.0000
1	0	0	-0.0025	0.0921	0.0958	0.9370	1.0000
0	0	0.0777	0.0020	0.0924	0.0936	0.9489	0.0511
0.5	0	0.0777	0.0028	0.0923	0.0931	0.9530	1.0000
1	0	0.0777	0.0023	0.0922	0.0946	0.9500	1.0000
0	0	0.157	0.0013	0.0920	0.0961	0.9360	0.0590
0.5	0	0.157	-0.0020	0.0919	0.0892	0.9620	0.9990
1	0	0.157	-0.0017	0.0920	0.0939	0.9449	1.0000
0	0.0777	0	0.0033	0.0922	0.0943	0.9359	0.0581
0.5	0.0777	0	-0.0023	0.0926	0.0922	0.9540	1.0000
1	0.0777	0	-0.0060	0.0926	0.0954	0.9519	1.0000
0	0.0777	0.0777	0.0039	0.0923	0.0954	0.9419	0.0561
0.5	0.0777	0.0777	0.0001	0.0924	0.0968	0.9459	0.9990
1	0.0777	0.0777	0.0034	0.0929	0.0928	0.9460	1.0000
0	0.0777	0.157	0.0007	0.0926	0.0934	0.9470	0.0470
0.5	0.0777	0.157	0.0035	0.0925	0.0945	0.9390	0.9990
1	0.0777	0.157	0.0022	0.0925	0.0959	0.9400	1.0000
0	0.157	0	-0.0036	0.0932	0.0990	0.9359	0.0601
0.5	0.157	0	0.0022	0.0935	0.0944	0.9459	1.0000
1	0.157	0	0.0055	0.0933	0.0987	0.9368	1.0000
0	0.157	0.0777	-0.0037	0.0935	0.0933	0.9520	0.0470
0.5	0.157	0.0777	0.0036	0.0932	0.0940	0.9479	0.9990
1	0.157	0.0777	-0.0038	0.0934	0.1007	0.9289	1.0000
0	0.157	0.157	-0.0022	0.0933	0.0975	0.9380	0.0600
0.5	0.157	0.157	0.0033	0.0937	0.0994	0.9360	1.0000
1	0.157	0.157	0.0000	0.0935	0.0974	0.9349	1.0000

Table 39: Monte Carlo simulation study results for fitting the step effect as fixed or random in a fixed cohort SWT design. These results are when the data were analyzed fitting the step effect as fixed.

			Fixed Step Model				
Theta	Sigma_subj	sigma_step	Bias	Model SE	MC SE	Coverage	Power/Type I Error Rate
0	0	0	0.0000	0.0918	0.0956	0.9419	0.0561
0.5	0	0	0.0026	0.0922	0.0958	0.9399	1.0000
1	0	0	-0.0025	0.0921	0.0958	0.9370	1.0000
0	0	0.0777	0.0021	0.0925	0.0935	0.9489	0.0501
0.5	0	0.0777	0.0029	0.0924	0.0931	0.9530	1.0000
1	0	0.0777	0.0023	0.0923	0.0946	0.9500	1.0000
0	0	0.157	0.0014	0.0923	0.0961	0.9369	0.0581
0.5	0	0.157	-0.0021	0.0921	0.0892	0.9620	0.9990
1	0	0.157	-0.0018	0.0923	0.0940	0.9450	1.0000
0	0.0777	0	0.0034	0.0922	0.0943	0.9359	0.0581
0.5	0.0777	0	-0.0023	0.0926	0.0922	0.9540	1.0000
1	0.0777	0	-0.0060	0.0926	0.0954	0.9520	1.0000
0	0.0777	0.0777	0.0041	0.0924	0.0954	0.9419	0.0551
0.5	0.0777	0.0777	0.0000	0.0925	0.0967	0.9460	0.9990
1	0.0777	0.0777	0.0034	0.0929	0.0928	0.9460	1.0000
0	0.0777	0.157	0.0007	0.0929	0.0933	0.9470	0.0470
0.5	0.0777	0.157	0.0035	0.0928	0.0945	0.9410	0.9990
1	0.0777	0.157	0.0021	0.0928	0.0959	0.9399	1.0000
0	0.157	0	-0.0036	0.0932	0.0989	0.9360	0.0600
0.5	0.157	0	0.0022	0.0935	0.0943	0.9460	1.0000
1	0.157	0	0.0053	0.0933	0.0988	0.9370	1.0000
0	0.157	0.0777	-0.0037	0.0936	0.0933	0.9530	0.0470
0.5	0.157	0.0777	0.0035	0.0933	0.0939	0.9489	0.9990
1	0.157	0.0777	-0.0039	0.0935	0.1006	0.9290	1.0000
0	0.157	0.157	-0.0023	0.0936	0.0975	0.9390	0.0600
0.5	0.157	0.157	0.0033	0.0940	0.0994	0.9360	1.0000
1	0.157	0.157	0.0001	0.0938	0.0974	0.9370	1.0000

Table 40: Simulation parameters for the Monte Carlo study on fitting the step effect as fixed or mean zero random effect in a fixed cohort SWT design.

Parameter	Variable	Values
Total number of clusters	I	10
Total number of steps	J	$I+1$
Subjects per cluster per step	K	10
Model intercept	μ	0.3
Effect of intervention	θ	$(0, 0.5, 1)$
Mean step effects	β	$\beta = (-0.2, -0.1, 0, 0.1, 0.2)$
i.e., Residual variance	σ^2	1.55^2
Random cluster effect variance	σ^2_{cl}	0.0777^2
Random subject effect variance	σ^2_{subj}	0.0777^2

In addition to analyzing these data with the correct model, we also analyzed them using a model that fits the step effects as mean zero random effects (See Equation 4.10).

$$y_{k(ij)} = (\mu + u_i + u_{k(i)}u_j) + \theta trt_{ij} + e_{k(ij)} \tag{4.10}$$

for cluster i, time step j, and subject k where

$$\begin{aligned} u_i &\sim N(0, \sigma^2_{cl}) \\ u_k(i) &\sim N(0, \sigma^2_{subj}) \\ u_j &\sim N(0, \sigma^2_{step}) \\ e_{k(ij)} &\sim N(0, \sigma^2) \end{aligned}$$

The simulations were run with 1000 Monte Carlo samples using the parameters outlined in Table (40).

For each analysis, the bias, mean model standard error, Monte Carlo standard error, coverage probability, and power/Type I error rate for the effect of intervention are reported in Table (41). As expected, when analyzing the data with the correct model using a fixed effect for

step, the estimation and inference for the intervention effect are both appropriate. However, when the data were analyzed using a mean zero random effect for step, there is bias in the estimation and the model-based standard error estimates are deflated compared to the Monte Carlo standard error. It appears that both the estimation and inference issues are actually worse for the smaller magnitude of the step slope (i.e., $|\beta| = 0.1$). When there was no true intervention effect (i.e., $\theta = 0$), these issues result in inflated Type I error rate when $\beta \neq 0$.

Lastly, we conducted a simulation study in which data were generated with an interaction term between intervention effect and step.

$$y_{k(ij)} = (\mu + u_i + u_{k(i)}) + \theta trt_{ij} + \beta t_j + \gamma t_j trt_{ij} + e_{k(ij)} \tag{4.11}$$

for cluster i, time step j, and subject k where

$$\begin{aligned} u_i &\sim N(0, \sigma^2_{cl}) \\ u_{k(i)} &\sim N(0, \sigma^2_{subj}) \\ e_{k(ij)} &\sim N(0, \sigma^2) \end{aligned}$$

In addition to analyzing these data with the correct model, we also analyzed them using a model that fits mean zero random effects for steps without including the interaction term between step and intervention (See Equation 4.12).

$$y_{k(ij)} = (\mu + u_i + u_{k(i)} + u_j) + \theta trt_{ij} + \beta t_j + e_{k(ij)} \tag{4.12}$$

for cluster i, time step j, and subject k where

$$\begin{aligned} u_i &\sim N(0, \sigma^2_{cl}) \\ u_{k(i)} &\sim N(0, \sigma^2_{subj}) \\ u_j &\sim N(0, \sigma^2_{step}) \\ e_{k(ij)} &\sim N(0, \sigma^2) \end{aligned}$$

The simulations were run with 1000 Monte Carlo samples using the parameters outlined in Table (42).

Table 41: Monte Carlo simulation study results for fitting the step effect as fixed or random in a fixed cohort SWT design.

Fixed Cohort SWT Design							
θ	β1	Step Effect	Bias	Model SE	MC SE	Coverage	Power
0	0	Random	-0.0024	0.0996	0.1023	0.9510	0.0470
0.5	0	Random	-0.0046	0.0993	0.0969	0.9620	0.9990
1	0	Random	-0.0009	0.0990	0.0989	0.9510	1.0000
0	0.1	Random	0.1679	0.1176	0.1515	0.6680	0.3270
0.5	0.1	Random	0.1633	0.1176	0.1574	0.6627	0.9960
1	0.1	Random	0.1583	0.1174	0.1526	0.6850	1.0000
0	0.2	Random	0.0933	0.1257	0.1374	0.8490	0.1450
0.5	0.2	Random	0.0910	0.1259	0.1375	0.8549	0.9960
1	0.2	Random	0.0726	0.1258	0.1365	0.8980	1.0000
0	-0.1	Random	-0.1621	0.1177	0.1477	0.6780	0.3100
0.5	-0.1	Random	-0.1676	0.1175	0.1561	0.6550	0.7360
1	-0.1	Random	-0.1624	0.1176	0.1470	0.6877	1.0000
0	-0.2	Random	-0.0809	0.1257	0.1398	0.8740	0.1210
0.5	-0.2	Random	-0.0890	0.1255	0.1371	0.8719	0.8779
1	-0.2	Random	-0.0880	0.1256	0.1410	0.8669	1.0000
0	0	Fixed	-0.0035	0.1283	0.1373	0.9319	0.0641
0.5	0	Fixed	-0.0085	0.1279	0.1266	0.9600	0.9700
1	0	Fixed	-0.0009	0.1280	0.1317	0.9419	1.0000
0	0.1	Fixed	0.0021	0.1285	0.1279	0.9500	0.0490
0.5	0.1	Fixed	0.0009	0.1281	0.1348	0.9340	0.9660
1	0.1	Fixed	-0.0048	0.1283	0.1306	0.9430	1.0000
0	0.2	Fixed	0.0076	0.1282	0.1309	0.9519	0.0471
0.5	0.2	Fixed	0.0054	0.1284	0.1305	0.9420	0.9780
1	0.2	Fixed	-0.0125	0.1282	0.1306	0.9380	1.0000
0	-0.1	Fixed	0.0006	0.1282	0.1271	0.9550	0.0430
0.5	-0.1	Fixed	-0.0021	0.1283	0.1343	0.9329	0.9739
1	-0.1	Fixed	0.0020	0.1285	0.1269	0.9510	1.0000
0	-0.2	Fixed	0.0044	0.1281	0.1323	0.9360	0.0620
0.5	-0.2	Fixed	-0.0037	0.1279	0.1294	0.9450	0.9640
1	-0.2	Fixed	-0.0029	0.1280	0.1345	0.9379	1.0000

The data were generated with a linear parametric term for step. The parameters θ and β represent the intervention effect and the linear step slope respectively. The 'Step Effect' column indicates whether the data were analyzed using a fixed effect (See Equation 4.9) or a mean zero random effect (See Equation 4.10). The bias, average model standard error, Monte Carlo standard error, coverage probability, and power are presented with respect to the effect of intervention parameter θ. Note that when $\theta = 0$, power actually refers to Type I error rate.

Table 42: Simulation parameters for the Monte Carlo study on fitting a random step effect in place of an interaction term between step and intervention effect in a fixed cohort SWT design.

Parameter	Variable	Values
Total number of clusters	I	10
Total number of steps	J	$I+1$
Subjects per cluster per step	K	10
Model intercept	μ	0.3
Effect of intervention	θ	$(0, 0.5, 1)$
Intervention and step interaction	γ	$(-0.05, -0.02, 0, 0.02, 0.05)$
Step slope parameter	β	$\beta = 0.25$
Residual variance	σ^2	1.55^2
Random cluster effect variance	σ^2_{cl}	0.0777^2
Random subject effect variance	σ^2_{subj}	0.0777^2

For each analysis, the bias, mean model standard error, Monte Carlo standard error, coverage probability, and power/Type I error rate for the effect of intervention are reported in Table (42). As expected, when analyzing the data with the correct model using a fixed effect for step, the estimation and inference for the intervention effect are both appropriate. When analyzing the data using the model with a random step effect, there is a small deflation of the standard error estimates and bias in the estimate for the main intervention effect. These issues result in inflation of Type I error rate even when the interaction term was zero. Also, for this analysis model, the coverage probability decreases for larger absolute values of the interaction term.

4.3.3 Expanding cohort SWT Designs

4.3.3.1 Choice of Correlation Structure Monte Carlo simulation studies were conducted to study the effect that choice of correlation structure (and misspecification) had

Table 43: This table summarizes the Monte Carlo simulation study results for fitting the step as a random effect in a fixed cohort SWT design with an underlying interaction term between step and intervention.

Fixed Cohort SWT							
θ	γ	Model	Bias	Model SE	MC SE	Coverage	Power
0	0	Random effect	0.0716	0.1267	0.1306	0.9109	0.0851
0.5	0	Random effect	0.0681	0.1266	0.1358	0.8950	0.9910
1	0	Random effect	0.0728	0.1263	0.1299	0.9080	1.0000
0	0.02	Random effect	0.1838	0.1267	0.1318	0.6697	0.3203
0.5	0.02	Random effect	0.1845	0.1270	0.1353	0.6777	1.0000
1	0.02	Random effect	0.1876	0.1263	0.1362	0.6830	1.0000
0	0.05	Random effect	0.3621	0.1271	0.1366	0.2130	0.7820
0.5	0.05	Random effect	0.3663	0.1267	0.1329	0.1870	1.0000
1	0.05	Random effect	0.3700	0.1271	0.1375	0.1982	1.0000
0	-0.02	Random effect	-0.0553	0.1265	0.1365	0.9100	0.0860
0.5	-0.02	Random effect	-0.0464	0.1265	0.1346	0.9110	0.9250
1	-0.02	Random effect	-0.0538	0.1264	0.1326	0.9209	1.0000
0	-0.05	Random effect	-0.2250	0.1263	0.1357	0.5560	0.4370
0.5	-0.05	Random effect	-0.2250	0.1265	0.1305	0.5626	0.5666
1	-0.05	Random effect	-0.2183	0.1264	0.1376	0.5832	1.0000
0	0	Interaction term	0.0019	0.2637	0.2770	0.9410	0.0580
0.5	0	Interaction term	0.0060	0.2639	0.2630	0.9490	0.4880
1	0	Interaction term	0.0076	0.2635	0.2656	0.9500	0.9690
0	0.02	Interaction term	0.0076	0.2637	0.2606	0.9520	0.0480
0.5	0.02	Interaction term	0.0036	0.2641	0.2647	0.9470	0.4780
1	0.02	Interaction term	-0.0089	0.2633	0.2726	0.9399	0.9609
0	0.05	Interaction term	-0.0006	0.2637	0.2636	0.9450	0.0550
0.5	0.05	Interaction term	-0.0034	0.2634	0.2677	0.9450	0.4660
1	0.05	Interaction term	0.0169	0.2639	0.2651	0.9530	0.9690
0	-0.02	Interaction term	-0.0037	0.2639	0.2719	0.9428	0.0572
0.5	-0.02	Interaction term	-0.0011	0.2635	0.2702	0.9399	0.4635
1	-0.02	Interaction term	-0.0122	0.2638	0.2697	0.9369	0.9570
0	-0.05	Interaction term	-0.0075	0.2639	0.2595	0.9480	0.0520
0.5	-0.05	Interaction term	-0.0041	0.2636	0.2722	0.9390	0.4630
1	-0.05	Interaction term	-0.0049	0.2638	0.2650	0.9410	0.9640

The data were generated with a linear parametric term for step. The parameters θ and γ represent the main intervention effect and the step by intervention interaction term respectively. The 'Model' column indicates whether the data were analyzed using the data generating model with an the interaction between step and intervention (See Equation 4.11) or a mean zero random effect (See Equation 4.12). The bias, average model standard error, Monte Carlo standard error, coverage probability, and power are presented with respect to the main effect of intervention parameter θ. Note that when $\theta = 0$, power actually refers to Type I error rate.

on the estimation and inference for the effect of intervention. Data were generated under the model shown in equation (4.13). Within-cluster correlation was modeled using random effects. Within-subject correlation was modeled through a covariance pattern with either an exchangeable or autoregressive structure.

$$y_{k(i)j} = \beta_0 + u_i + u_{k(i)} + \beta_1 trt_{ij} + \beta_2 t_j + e_{k(i)j} \tag{4.13}$$

for cluster i, time step j, and subject k where $u_i \sim N(0, \sigma^2_{cl})$ are the random effects for clusters and $u_{k(i)} \sim N(0, \sigma^2_{subj})$ are the random effects for subjects.

$$e_{k(i)j} \sim N(0, \sigma^2)$$

The parameters for the simulation were chosen as shown in Table (44).

To simulate subject enrollment, a multinomial random variable was generated for each subject with probability of enrolling during the first step equaling 0.6 and the probability of enrolling during step j equal to $\frac{0.4}{J-1}$.

Table 44: Simulation parameters for the Monte Carlo study on choice of correlation structure for an expanding cohort SWT designs.

Parameter	Variable	Values
Total number of clusters	I	10
Total number of steps	J	$I+1$
Subjects per cluster per step	K	20
Model intercept	β_0	0.3
Effect of intervention	β_1	$(0, 0.5)$
Step Effect	β_2	0.25
Residual variance	σ^2	1.55^2
Correlation parameter	ρ	0.6

Each simulated data set will be analyzed using linear mixed models with that fit random effects for the clusters and use a covariance pattern to handle within-subject correlation. Since expanding cohort designs include unbalanced outcome due to late recruitment, GEE was not be used as an analysis. For each analysis, the bias, mean model-based standard error, Monte Carlo standard error, coverage probability, and power/Type I error rate for the effect of intervention are reported in Table (45).

Table (45) demonstrates that when the correct correlation structure is chosen for a LMM, the bias, standard error estimation, coverage, and power exhibit good properties. When autoregressive generated data are analyzed as exchangeable, the standard error estimate is deflated compared to the Monte Carlo standard error. Converesly, when the data are generated as exchangeable, but are analyzed as autoregressive, the standard error estimate is inflated compared to the Monte Carlo standard error. In terms of Type I error rate, this results in inflated Type I error rate when the true underlying correlation structure is autoregressive, but fit as exchangeable and deflated (conservative) Type I error rate when the true underlying correlation structure is exchangeable, but fit as autoregressive.

Table 45: Results from the Monte Carlo simulation study for an expanding cohort SWT design.

θ	Correlation Generated	Analysis Correlation	Bias	MC Standard Error	Estimated Standard Error	Coverage	Power
0.5	AR	AR	0.0009	0.1483	0.1455	0.9410	0.9360
0.5	AR	EXCH	-0.0005	0.1860	0.1408	0.8650	0.8860
0	AR	AR	-0.0013	0.1440	0.1457	0.9460	0.0540
0	AR	EXCH	0.0051	0.1888	0.1408	0.8620	0.1380
0.5	EXCH	AR	-0.0001	0.1671	0.1834	0.9730	0.7900
0.5	EXCH	EXCH	0.0015	0.1317	0.1311	0.9390	0.9640
0	EXCH	AR	-0.0025	0.1686	0.1837	0.9710	0.0290
0	EXCH	EXCH	-0.0019	0.1289	0.1313	0.9580	0.0420

Theta is the intervention effect parameter. The abbreviations 'EXCH' and 'AR' stand for exchangeable and autoregressive respectively. The MC error represents the Monte Carlo standard error. Note that when theta is zero, 'Power' refers to Type I error rate.

4.3.3.2 Fixed versus Random Effects for Steps Monte Carlo simulations were conducted to study the effect of fitting the step effect as fixed versus random had on estimation and inference for the effect of the intervention. Data was generated according to the model outlined in the equation below.

$$y_{k(ij)} = (\mu + u_i + u_j + u_{k(i)}) + \beta_1 trt_{ij} + \beta_j + e_{k(ij)} \tag{4.14}$$

for cluster i, time step j, and subject k where

$$u_i \sim N(0, \sigma^2_{cl})$$
$$u_j \sim N(0, \sigma^2_{step})$$
$$u_{k(i)} \sim N(0, \sigma^2_{subj})$$
$$e_{k(ij)} \sim N(0, \sigma^2)$$

To simulate subject enrollment , a multinomial random variable was generated for each subject with probability of enrolling during the first step equaling 0.6 and the probability of enrolling during step j equal to $\frac{0.4}{J-1}$.

Table 46: Simulation parameters for the Monte Carlo study on fitting the step effect as fixed or random in an expanding cohort SWT design.

Parameter	Variable	Values
Total number of clusters	I	10
Total number of steps	J	$I+1$
Subjects per cluster	K	20
Model intercept	μ	0.3
Effect of intervention	β_1	$(0, 0.5, 1)$
Mean step effects	β_j	$\beta_j = (0.25(j-1))$ for $j = 2, \ldots, J$
Residual variance	σ^2	1.55^2
Random cluster effect variance	σ^2_{cl}	0.0777^2
Random step effect variance	σ^2_{step}	$(0, 0.0777^2, 0.157^2)$
Random subject effect variance	σ^2_{subj}	0.1

All simulations shared the parameters outlined in Table 46. Note that the mean step effects were chosen to fit a linear trend.

One thousand Monte Carlo samples were generated for each choice of parameters. For each sample, the data was analyzed both by fitting the step effect as random and by only fitting the step effect as fixed. For each analysis, the bias, mean model standard error, Monte Carlo standard error, coverage probability, and power/Type I error rate for the effect of intervention are reported in Tables (47) and (48). Simulation results do not suggest any differences in estimation or inference for the treatment effect based on fitting the step effect as random or fixed.

For the second random step simulation study, we generated data according to equation (4.15) below.

$$y_{k(i)j} = (\mu + u_i + u_k(i)) + \theta trt_{ij} + \beta t_j + e_{k(i)j} \tag{4.15}$$

for cluster i, time step j, and subject k

Table 47: Monte Carlo simulation study results for fitting the step effect as fixed or random in an expanding cohort SWT design. These results are when the data were analyzed fitting the step effect as random.

			Random Step Model				
Theta	Sigma_subj	sigma_step	Bias	Model SE	MC SE	Coverage	Power/Type I Error Rate
0	0	0	0.0006	0.1026	0.1047	0.9520	0.0480
0.5	0	0	0.0014	0.1027	0.1027	0.9429	0.9980
1	0	0	-0.0013	0.1020	0.1056	0.9429	1.0000
0	0	0.0777	-0.0018	0.1022	0.1022	0.9539	0.0441
0.5	0	0.0777	0.0070	0.1024	0.1054	0.9340	1.0000
1	0	0.0777	0.0035	0.1020	0.1056	0.9439	1.0000
0	0	0.157	0.0015	0.1023	0.1046	0.9460	0.0520
0.5	0	0.157	-0.0013	0.1021	0.1039	0.9499	0.9970
1	0	0.157	0.0066	0.1026	0.1076	0.9380	1.0000
0	0.0777	0	0.0008	0.1023	0.1054	0.9358	0.0622
0.5	0.0777	0	0.0086	0.1026	0.1072	0.9320	0.9980
1	0.0777	0	-0.0003	0.1029	0.1059	0.9410	1.0000
0	0.0777	0.0777	-0.0057	0.1026	0.1023	0.9440	0.0520
0.5	0.0777	0.0777	-0.0021	0.1029	0.1065	0.9470	0.9990
1	0.0777	0.0777	-0.0028	0.1028	0.1041	0.9460	1.0000
0	0.0777	0.157	0.0042	0.1028	0.1075	0.9400	0.0540
0.5	0.0777	0.157	0.0034	0.1028	0.1070	0.9409	0.9960
1	0.0777	0.157	-0.0048	0.1026	0.1048	0.9469	1.0000
0	0.157	0	0.0013	0.1035	0.1051	0.9429	0.0521
0.5	0.157	0	0.0013	0.1037	0.1086	0.9419	0.9960
1	0.157	0	0.0097	0.1041	0.1090	0.9359	1.0000
0	0.157	0.0777	0.0039	0.1040	0.1057	0.9439	0.0531
0.5	0.157	0.0777	-0.0028	0.1037	0.1083	0.9330	0.9990
1	0.157	0.0777	0.0042	0.1038	0.1043	0.9480	1.0000
0	0.157	0.157	0.0014	0.1039	0.1065	0.9449	0.0541
0.5	0.157	0.157	-0.0060	0.1040	0.1118	0.9300	0.9950
1	0.157	0.157	0.0067	0.1039	0.1029	0.9510	1.0000

Table 48: Monte Carlo simulation study results for fitting the step effect as fixed or random in an expanding cohort SWT design. These results are when the data were analyzed fitting the step effect as fixed.

			Fixed Step Model				
Theta	Sigma_subj	sigma_step	Bias	Model SE	MC SE	Coverage	Power/Type I Error Rate
0	0	0	0.0007	0.1026	0.1048	0.9520	0.0480
0.5	0	0	0.0016	0.1027	0.1027	0.9440	0.9980
1	0	0	-0.0013	0.1020	0.1056	0.9420	1.0000
0	0	0.0777	-0.0017	0.1023	0.1022	0.9560	0.0410
0.5	0	0.0777	0.0070	0.1025	0.1054	0.9350	1.0000
1	0	0.0777	0.0036	0.1021	0.1055	0.9440	1.0000
0	0	0.157	0.0014	0.1026	0.1045	0.9460	0.0520
0.5	0	0.157	-0.0012	0.1024	0.1039	0.9510	0.9970
1	0	0.157	0.0066	0.1029	0.1077	0.9390	1.0000
0	0.0777	0	0.0007	0.1023	0.1057	0.9350	0.0630
0.5	0.0777	0	0.0088	0.1027	0.1071	0.9319	0.9980
1	0.0777	0	-0.0003	0.1029	0.1058	0.9410	1.0000
0	0.0777	0.0777	-0.0056	0.1027	0.1024	0.9439	0.0521
0.5	0.0777	0.0777	-0.0021	0.1029	0.1065	0.9470	0.9990
1	0.0777	0.0777	-0.0028	0.1029	0.1041	0.9480	1.0000
0	0.0777	0.157	0.0042	0.1031	0.1075	0.9410	0.0540
0.5	0.0777	0.157	0.0032	0.1031	0.1071	0.9390	0.9960
1	0.0777	0.157	-0.0047	0.1029	0.1050	0.9470	1.0000
0	0.157	0	0.0013	0.1036	0.1050	0.9430	0.0520
0.5	0.157	0	0.0014	0.1037	0.1085	0.9420	0.9960
1	0.157	0	0.0096	0.1042	0.1089	0.9360	1.0000
0	0.157	0.0777	0.0037	0.1041	0.1058	0.9450	0.0520
0.5	0.157	0.0777	-0.0028	0.1038	0.1083	0.9339	0.9990
1	0.157	0.0777	0.0042	0.1039	0.1044	0.9479	1.0000
0	0.157	0.157	0.0012	0.1041	0.1066	0.9470	0.0520
0.5	0.157	0.157	-0.0061	0.1043	0.1119	0.9299	0.9950
1	0.157	0.157	0.0066	0.1042	0.1027	0.9520	1.0000

$$u_i \sim N(0, \sigma^2_{cl})$$
$$u_k(i) \sim N(0, \sigma^2_{subj})$$
$$e_{k(i)j} \sim N(0, \sigma^2)$$

In addition to analyzing these data with the correct model, we also analyzed them using a model that fits the step effects as mean zero random effects (See Equation 4.16).

$$y_{k(ij)} = (\mu + u_i + u_{k(i)}u_j) + \theta trt_{ij} + e_{k(ij)} \tag{4.16}$$

for cluster i, time step j, and subject k where

$$u_i \sim N(0, \sigma^2_{cl})$$
$$u_k(i) \sim N(0, \sigma^2_{subj})$$
$$u_j \sim N(0, \sigma^2_{step})$$
$$e_{k(ij)} \sim N(0, \sigma^2)$$

The simulations were run with 1000 Monte Carlo samples using the parameters outlined in Table (49). To simulate subject enrollment , a multinomial random variable was generated for each subject with probability of enrolling during the first step equaling 0.6 and the probability of enrolling during step j equal to $\frac{0.4}{J-1}$.

For each analysis, the bias, mean model standard error, Monte Carlo standard error, coverage probability, and power/Type I error rate for the intervention effect are reported in Table (50). As expected, when analyzing the data with the correct model using a fixed effect for step, the estimation and inference for the intervention effect are both appropriate. However, when the data were analyzed using a mean zero random effect for step, there is bias in the estimation and the model-based standard error estimates are deflated compared to the Monte Carlo standard error. It appears that both the estimation and inference issues are actually worse for the smaller magnitude of the step slope (i.e., $|\beta| = 0.1$). When there was no true intervention effect (i.e., $\theta = 0$), these issues result in inflated Type I error rate when $\beta \neq 0$.

Table 49: Simulation parameters for the Monte Carlo study on fitting the step effect as fixed or mean zero random effect in an expanding cohort SWT design.

Parameter	Variable	Values
Total number of clusters	I	10
Total number of steps	J	$I+1$
Subjects per cluster per step	K	10
Model intercept	μ	0.3
Effect of intervention	θ	$(0, 0.5, 1)$
Mean step effects	β	$\beta = (-0.2, -0.1, 0, 0.1, 0.2)$
i.e., Residual variance	σ^2	1.55^2
Random cluster effect variance	σ^2_{cl}	0.0777^2
Random subject effect variance	σ^2_{subj}	0.0777^2

Lastly, we conducted a simulation study in which data were generated with an interaction term between intervention effect and step.

$$y_{k(ij)} = (\mu + u_i + u_{k(i)}) + \theta trt_{ij} + \beta t_j + \gamma t_j trt_{ij} + e_{k(ij)} \tag{4.17}$$

for cluster i, time step j, and subject k where

$$u_i \sim N(0, \sigma^2_{cl})$$
$$u_{k(i)} \sim N(0, \sigma^2_{subj})$$
$$e_{k(ij)} \sim N(0, \sigma^2)$$

In addition to analyzing these data with the correct model, we also analyzed them using a model that fits mean zero random effects for steps without including the interaction term between step and intervention (See Equation 4.18).

Table 50: Monte Carlo simulation study results for fitting the step effect as fixed or random in a fixed cohort SWT design.

Expanding Cohort SWT Design							
θ	**β1**	**Step Effect**	**Bias**	**Model SE**	**MC SE**	**Coverage**	**Power**
0	0	Random	0.0057	0.1112	0.1059	0.9640	0.0350
0.5	0	Random	-0.0065	0.1112	0.1135	0.9510	0.9910
1	0	Random	0.0012	0.1115	0.1104	0.9470	1.0000
0	0.1	Random	0.1975	0.1284	0.1668	0.6400	0.3530
0.5	0.1	Random	0.1862	0.1286	0.1621	0.6527	0.9970
1	0.1	Random	0.1909	0.1286	0.1643	0.6470	1.0000
0	0.2	Random	0.1025	0.1392	0.1494	0.8659	0.1261
0.5	0.2	Random	0.1043	0.1393	0.1572	0.8590	0.9820
1	0.2	Random	0.1085	0.1397	0.1552	0.8550	1.0000
0	-0.1	Random	-0.2015	0.1279	0.1677	0.6180	0.3720
0.5	-0.1	Random	-0.1970	0.1281	0.1647	0.6270	0.6180
1	-0.1	Random	-0.1854	0.1283	0.1662	0.6650	0.9990
0	-0.2	Random	-0.1015	0.1394	0.1532	0.8700	0.1280
0.5	-0.2	Random	-0.1123	0.1396	0.1495	0.8507	0.7786
1	-0.2	Random	-0.1021	0.1396	0.1544	0.8570	1.0000
0	0	Fixed	0.0043	0.1428	0.1398	0.9530	0.0460
0.5	0	Fixed	-0.0057	0.1428	0.1461	0.9390	0.9270
1	0	Fixed	0.0013	0.1427	0.1454	0.9490	1.0000
0	0.1	Fixed	0.0034	0.1430	0.1438	0.9489	0.0480
0.5	0.1	Fixed	-0.0041	0.1426	0.1406	0.9580	0.9310
1	0.1	Fixed	0.0008	0.1428	0.1449	0.9430	1.0000
0	0.2	Fixed	-0.0035	0.1424	0.1395	0.9590	0.0390
0.5	0.2	Fixed	-0.0026	0.1427	0.1468	0.9370	0.9200
1	0.2	Fixed	0.0004	0.1431	0.1451	0.9530	1.0000
0	-0.1	Fixed	-0.0077	0.1426	0.1443	0.9470	0.0490
0.5	-0.1	Fixed	-0.0042	0.1425	0.1436	0.9540	0.9280
1	-0.1	Fixed	0.0064	0.1426	0.1423	0.9559	1.0000
0	-0.2	Fixed	0.0044	0.1427	0.1428	0.9440	0.0510
0.5	-0.2	Fixed	-0.0048	0.1429	0.1400	0.9619	0.9339
1	-0.2	Fixed	0.0041	0.1429	0.1455	0.9380	1.0000

Table 51: Simulation parameters for the Monte Carlo study on fitting a random step effect in place of an interaction term between step and intervention effect in a fixed cohort SWT design.

Parameter	Variable	Values
Total number of clusters	I	10
Total number of steps	J	$I+1$
Subjects per cluster per step	K	10
Model intercept	μ	0.3
Effect of intervention	θ	$(0, 0.5, 1)$
Intervention and step interaction	γ	$(-0.05, -0.02, 0, 0.02, 0.05)$
Step slope parameter	β	$\beta = 0.25$
Residual variance	σ^2	1.55^2
Random cluster effect variance	σ^2_{cl}	0.0777^2
Random subject effect variance	σ^2_{subj}	0.0777^2

The data were generated with a model shown in Equation (4.18).

$$y_{k(ij)} = (\mu + u_i + u_{k(i)} + u_j) + \theta trt_{ij} + \beta t_j + e_{k(ij)} \tag{4.18}$$

for cluster i, time step j, and subject k where

$$u_i \sim N(0, \sigma^2_{cl})$$
$$u_{k(i)} \sim N(0, \sigma^2_{subj})$$
$$u_j \sim N(0, \sigma^2_{step})$$
$$e_{k(ij)} \sim N(0, \sigma^2)$$

The simulations were run with 1000 Monte Carlo samples using the parameters outlined in Table (51). To simulate subject enrollment , a multinomial random variable was generated for each subject with probability of enrolling during the first step equaling 0.6 and the probability of enrolling during step j equal to $\frac{0.4}{J-1}$.

For each analysis, the bias, mean model standard error, Monte Carlo standard error, coverage probability, and power/Type I error rate for the effect of intervention are reported in Table

(52). As expected, when analyzing the data with the correct model using a fixed effect for step, the estimation and inference for the intervention effect are both appropriate. When analyzing the data using the model with a random step effect, there is a small deflation of the standard error estimates and bias in the estimate for the main intervention effect. These issues result in inflation of Type I error rate even when the interaction term was zero. Also, for this analysis model, the coverage probability decreases for larger absolute values of the interaction term.

4.4 DISCUSSION

When choosing the correlation structure when using a LMM, it is clear that incorrectly choosing the correlation structure results in improper inference for the effect of intervention. Interestingly, choosing an exchangeable correlation structure when the underlying structure is autoregressive results in underestimation of the standard error and inflated Type I error rate. Conversely, choosing an autoregressive correlation structure when the underlying structure is exchangeable results in overestimation of the standard error and conservative Type I error rate. Thus, when using LMMs to analyze SWT data, it is advisable to test choice of correlation structure through the use information criteria such as AIC[1] or BIC[17]. When using GEE for SWT data, the empirical variance estimator underestimates the standard error. This effect is likely due to reliance on asymptotic results. Future work on this subject could include both varying the total number of clusters and subjects per cluster. Additionally, finite sample size corrected estimators for GEE should be studied. While all GEE-based methods resulted in underestimated standard error, our study suggests that choosing the correlation unit as clusters rather than the subjects nested within clusters yield less underestimated standard error.

When fitting the step effect, it appears that whether it is included as a fixed or random effect does not substantially affect the estimation or inference for the effect of intervention. Thus, the decision of whether to fit the step effect as fixed or random may be chosen depending on

Table 52: This table summarizes the Monte Carlo simulation study results for fitting the step as a random effect in a fixed cohort SWT design with an underlying interaction term between step and intervention.

Cross-sectional SWT							
θ	γ	Model	Bias	Model SE	MC SE	Coverage	Power
0	0	Random effect	0.0720	0.1258	0.1319	0.9080	0.0900
0.5	0	Random effect	0.0729	0.1256	0.1345	0.8980	0.9920
1	0	Random effect	0.0608	0.1256	0.1331	0.9090	1.0000
0	0.02	Random effect	0.1807	0.1260	0.1317	0.6810	0.3100
0.5	0.02	Random effect	0.1840	0.1255	0.1328	0.6800	0.9990
1	0.02	Random effect	0.1855	0.1260	0.1318	0.6750	1.0000
0	0.05	Random effect	0.3724	0.1261	0.1419	0.1900	0.8040
0.5	0.05	Random effect	0.3674	0.1263	0.1341	0.1710	1.0000
1	0.05	Random effect	0.3588	0.1264	0.1367	0.2160	1.0000
0	-0.02	Random effect	-0.0503	0.1253	0.1364	0.9020	0.0970
0.5	-0.02	Random effect	-0.0456	0.1253	0.1327	0.9190	0.9300
1	-0.02	Random effect	-0.0484	0.1257	0.1353	0.9090	1.0000
0	-0.05	Random effect	-0.2250	0.1257	0.1346	0.5460	0.4490
0.5	-0.05	Random effect	-0.2320	0.1256	0.1318	0.5290	0.5380
1	-0.05	Random effect	-0.2275	0.1257	0.1331	0.5530	1.0000
0	0	Interaction term	0.0087	0.2634	0.2735	0.9330	0.0670
0.5	0	Interaction term	-0.0028	0.2632	0.2607	0.9570	0.4640
1	0	Interaction term	-0.0097	0.2632	0.2602	0.9450	0.9600
0	0.02	Interaction term	0.0017	0.2633	0.2665	0.9520	0.0480
0.5	0.02	Interaction term	-0.0090	0.2631	0.2655	0.9449	0.4635
1	0.02	Interaction term	-0.0060	0.2632	0.2622	0.9540	0.9690
0	0.05	Interaction term	0.0130	0.2632	0.2676	0.9419	0.0561
0.5	0.05	Interaction term	0.0035	0.2635	0.2690	0.9370	0.4800
1	0.05	Interaction term	0.0137	0.2637	0.2635	0.9510	0.9720
0	-0.02	Interaction term	0.0001	0.2629	0.2742	0.9440	0.0560
0.5	-0.02	Interaction term	0.0194	0.2632	0.2749	0.9350	0.5150
1	-0.02	Interaction term	-0.0001	0.2634	0.2558	0.9540	0.9680
0	-0.05	Interaction term	0.0045	0.2634	0.2700	0.9510	0.0490
0.5	-0.05	Interaction term	-0.0016	0.2632	0.2694	0.9420	0.4750
1	-0.05	Interaction term	-0.0061	0.2635	0.2587	0.9560	0.9670

The data were generated with a linear parametric term for step. The parameters θ and γ represent the main intervention effect and the step by intervention interaction term respectively. The 'Model' column indicates whether the data were analyzed using the data generating model with an the interaction between step and intervention (See Equation 4.17) or a mean zero random effect (See Equation 4.18). The bias, average model standard error, Monte Carlo standard error, coverage probability, and power are presented with respect to the main effect of intervention parameter θ. Note that when $\theta = 0$, power actually refers to Type I error rate.

circumstances of the trial such as how many steps are in the study. It should be noted that when we the step effect as random, we still included fixed effect terms so that the mean for these effects was not zero. This detail is important so that background temporal effects can be controlled.

5.0 DISCUSSION

From the review work in Chapter 2, it is clear that reporting of important design and analytic features of SWTs is greatly lacking. To classify a SWT as one of the three types that I proposed (i.e., fixed cohort, expanding cohort, and cross-sectional), recruitment and schedule of outcome assessment must be clearly reported. While many of the trials in the systematic review could be classified into one of the three types, in many cases the recruitment and outcome assessment features were not directly stated, but had to be inferred. Additionally, without clearing specifying the basic design and analytical features of the SWT, it is unclear how accurate reported power and sample size estimations are. These deficiencies in reporting could be corrected by adhering to the additional CONSORT guidelines for SWT designs as recommended by Hemming et al.[7]

In Chapter 3, we examined several design assumptions through Monte Carlo simulation studies. These design features included homogeneity/heterogeneity of temporal trends, time-varying/time-fixed intervenion effect, and cluster-level versus subject-level temporal trends. Our simulation study of heterogeneous temporal trends indicate the importance of accounting for such heterogeneity across different clusters. When the analysis model assumes homogeneous temporal trends in the presence of heterogeneous trends, the result is improper estimation of the standard errors for the intervention effect parameter. This issue can lead to inflated Type I error rate. For the time-varying effect of intervention study, we found that an analysis assuming a time-fixed intervention effect will produce biased estimates for the main intervention effect.

Lastly, we conducted a simulation study that examined cluster-level and subject-level trends in expanding cohort trials. Our simulation study concluded that including the cluster-level trend is necessary for proper estimation of the intervention effect. On the other hand, omission of subject-level trends does not appear to have a negative effect on estimation or inference for the effect of intervention.

In Chapter 4, we studied several analytical issues for SWT designs through Monte Carlo simulations. Specifically, we studied choice of correlation structure and fitting the step effect as a random effect. For choice of correlation structure, we found that fitting an exchangeable structure when the underlying data were first order autoregressive resulted in deflated standard error estimates and inflated Type I error rate. Conversely, fitting the first-order autoregressive structure when the underlying data were exchangeable results in inflated standard error estimates and overly conservative Type I error rate. These observations are especially important when one notes from Chapter 2 that most SWTs are analyzed using random effects, which for Gaussian error terms is equivalent to an exchangeable correlation structure. When studying fitting step effects as random effects, we conducted three different studies: zero mean random effects, non-zero mean random effects, and random effects when an interaction is present. For zero mean random effects, we found that fitting such effects when the underlying data had fixed effect terms resulted in biased estimation for the intervention effect. These results make sense in that the fixed effects were accounted for a temporal trend in the mean outcome. The study on non-zero mean random effects, conversely, found that choice of fitting the step effect as fixed or random did not appreciably affect estimation or inference for the intervention effect. Lastly, for the study on using random step effects in the presence of an interaction term, we found that the random effect analysis yielded substantial bias in the estimation of the intervention effect.

Based upon the simulation studies conducted in Chapters 3 and 4, I believe that additional recommendations are needed for reporting guidelines for SWTs. While the recommendations by Hemming et al.[7] ensure that the fundamental design issues and final analysis model are reported, I suggest there should be standard reporting for exploration of assumptions and model choices. The simulation studies I have conducted clearly demonstrate that there is

potential for bias or improper statistical inference for the intervention effect when the analysis model is not correctly specified. There should be reporting of exploratory analyses regarding heterogeneity of temporal trends, time-varying treatment effect, and choice of correlation structure.

BIBLIOGRAPHY

[1] H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6), 1974.

[2] G. Baio, A. Copas, G. Ambler, J. Hargreaves, E. Beard, and R. Omar. Sample size calculation for a stepped wedge trial. *Trials*, 16(354), 2014.

[3] P. Byass. The global burden of liver disease: a challenge for methods and for public health. *BMC Medicine*, 12(159), 2014.

[4] A. Copas, J. Lewis, J. Thompson, C. Davey, G. Baio, and J. Hargreaves. Designing a stepped wedge trial: three main designs, carry-over effects and randomisation approaches. *Trials*, 16(352), 2014.

[5] M. Grayling, J. Wason, and A. Mander. Stepped wedge cluster randomized controlled trial design: A review of reporting quality and design features. *Trials*, 18(33), 2017.

[6] K. Hemming and A. Girling. Stepped-wedge cluster randomised controlled trials: some variations on the common design. *Trials*, 14(Suppl 1):P15, 2013.

[7] K. Hemming and T. Haines. The stepped wedge cluster randomised trial: rationale, design, analysis, and reporting. *BMJ*, 2015.

[8] R. Hooper, S. Teerenstra, E. de Hoop, and S. Eldridge. Sample size calculation for stepped wedge and other longitudinal cluster randomised trials. *Statistics in Medicine*, 35(26), 2016.

[9] J. Hughes, T. Granston, and P. Heagerty. Current issues in the design and analysis of stepped wedge trials. *Contemporary Clinical Trials*, 45(Pt A):55–60, 2015.

[10] M. Hussey and J. Hughes. Design and analysis of stepped wedge cluster randomized trials. *Contemporary Clinical Trials"*, 28, 2007.

[11] SAS Institute Inc. The glimmix procedure: Fit statistics.

[12] J. Kasca, K. Hemming, J.N.S. Matthews, and A. Forbes. Impact of non-uniform correlation structure on sample size and power in multiple-period cluster randomised trials. *Sage Journals*, 2017.

[13] K. Liang and S. Zeger. Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13, 1986.

[14] P. Marcellin and B. Kutala. Liver diseases: A major, neglected global public health problem requiring urgent actions and large-scale screening. *Liver International*, 38, 2018.

[15] S. Nazzani and F. Preissr. In-hospital length of stay after major surgical oncological procedures. *European Journal of Surgical Oncology*, 2018.

[16] W. Pan. Akaike's information criterion in generalized estimating equations. *Biometrics*, 57(1), 2004.

[17] G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2), 1978.

[18] J. Scott, A. deCamp, M. Juraska, M. Fay, and P. Gilbert. Finite-sample corrected generalized estimating equation of population average treatment effects in stepped wedge cluster randomized trials. *Stat Methods Med Res*, 2014.

[19] W. Woertman, E. Hoop, M. Moerbeek, S. Zuidema, D. Gerritsen, and S. Teerenstra. Stepped wedge designs could reduce the required sample size in cluster randomized trials. *J Clin Epidemiol*, 66(7):752–758, 2013.